Universidade de Lisboa

Faculdade de Ciências
Departamento de Estatística e Investigação Operacional

Análise de sobrevivência com acontecimentos múltiplos: Aplicação ao estudo do tempo até à ocorrência de enfarte do miocárdio

Adriana Branco de Andrade e Belo Cabete

Dissertação
Mestrado em Bioestatística

2012

Universidade de Lisboa

Faculdade de Ciências
Departamento de Estatística e Investigação Operacional

Análise de sobrevivência com acontecimentos múltiplos: Aplicação ao estudo do tempo até à ocorrência de enfarte do miocárdio

Adriana Branco de Andrade e Belo Cabete

Dissertação orientada pela
Professora Doutora Cristina Maria Tristão Simões Rocha

Mestrado em Bioestatística

2012

Resumo

O enfarte agudo do miocárdio constitui atualmente uma das principais causas de morte em todo o mundo. Este trabalho tem como objetivo a modelação do tempo até à ocorrência de múltiplos enfartes como complicações de uma síndrome coronária aguda e a determinação dos seus fatores de risco de entre as variáveis registadas na admissão hospitalar.

Podendo existir várias ocorrências do mesmo acontecimento para o mesmo indivíduo, a utilização do modelo de Cox para tempos independentes não é adequada. De entre os diversos modelos de sobrevivência para acontecimentos múltiplos, optou-se por considerar o modelo de regressão PWP desenvolvido por Prentice, Williams e Peterson (1981) para a avaliação da influência dos diversos fatores no tempo de vida dos indivíduos. Este modelo revelou-se o mais adequado nesta situação de acontecimentos ordenados com risco condicional, uma vez que o risco de sofrer cada novo enfarte é diferente do risco associado ao enfarte anterior, sendo assumido que o doente apenas está em risco de sofrer o enfarte de ordem k quando tiver sofrido o enfarte de ordem k-1. O modelo será aplicado a dados reais respeitantes a um estudo envolvendo doentes hospitalizados com síndrome coronária aguda na Unidade de Cuidados Intensivos de Cardiologia dos Hospitais da Universidade de Coimbra.

Palavras-chave: análise de sobrevivência, modelo PWP, acontecimentos múltiplos, acontecimentos recorrentes, estimador robusto da matriz de covariância, enfarte agudo do miocárdio.

Abstract

Nowadays, acute myocardial infarction (AMI) is a major cause of death in developed countries. The aim of this work is to study the time to recurrent infarctions as complications of an acute coronary syndrome and to determine their risk factors among information collected at patient's admission in the hospital.

Since, in this case, the event of interest may occur several times for the same individual, the use of Cox proportional hazards model is not appropriate. Among the various multiple event time models, it was considered the PWP regression model proposed by Prentice, Williams and Peterson (1981) to evaluate the influence of different covariates in survival. This model proved to be the most suitable for ordered multiple events with conditional risk, whereas the risk of each new AMI is different from the risk of the previous one, being assumed that a patient can only be at risk for the kth event after experiencing the (k-1)th event. The model will be applied to real data from a study involving hospitalized patients with acute coronary syndrome in the Cardiology Intensive Care Unit of the University Hospital of Coimbra.

Keywords: survival analysis, PWP model, multiple events, recurrent events, robust covariance matrix estimator, acute myocardial infarction.

Agradecimentos

Uma tese de mestrado, apesar de ser fruto de um trabalho árduo e solitário, reúne o contributo de várias pessoas. Desde que me propus enfrentar este desafio, contei com a confiança e o apoio de inúmeras pessoas e instituições, sem as quais este trabalho não teria sido possível.

À Professora Doutora Cristina Rocha, orientadora da dissertação, agradeço toda a partilha de conhecimentos. Agradeço ainda a compreensão e os momentos de boa disposição que me ajudaram a superar os períodos mais difíceis.

À Direção da Sociedade Portuguesa de Cardiologia, em particular ao Professor Doutor Lino Gonçalves, por me ter dado a possibilidade de conciliar o meu trabalho com as aulas e a realização desta dissertação.

Ao Professor Doutor Pedro Monteiro e a todos os profissionais do serviço de cardiologia dos Hospitais da Universidade de Coimbra, por tão gentilmente terem cedido os dados, fruto do seu árduo trabalho diário, sem os quais não teria sido possível a realização desta dissertação.

A todos os meus amigos e colegas que sempre estiveram presentes incentivando-me com carinho.

Aos meus familiares pelo incentivo recebido ao longo destes meses, pelo amor, alegria e atenção sem reservas...

"Só sabemos com exatidão quando sabemos pouco,
à medida que vamos adquirindo conhecimentos,
instala-se a dúvida."
Johann Goethe

Lista de siglas

AD: aurícula direita.

AE: aurícula esquerda.

AI: angina instável.

AIT: acidente isquémico transitório.

AVC: acidente vascular cerebral.

BB: beta-bloqueantes.

CABG: cirurgia de *bypass*.

Classe kk: classe Killip-Kimball.

CT: colesterol total.

DAP: doença arterial periférica.

DM: diabetes *mellitus*.

Diag: diagnóstico.

EAM: enfarte agudo do miocárdio.

ECG: eletrocardiograma.

FC: frequência cardíaca.

Glic: glicémia.

HDL: colesterol HDL.

HTA: hipertensão arterial.

ICC: insuficiência cardíaca.

IECA: inibidor da enzima da conversão da angiotensina.

IMC: índice de massa corporal.

LDL: colesterol LDL.

PTCA: angioplastia coronária.

TA: tensão arterial.

TAD: tensão arterial diastólica.

TAS: tensão arterial sistólica.

TFG: taxa de filtração glomerular.

TG: triglicerídeos.

VD: ventrículo direito.

VE: ventrículo esquerdo.

Índice

Nota Introdutória

A modelação do tempo de vida, compreendido entre um instante inicial e a ocorrência de um único acontecimento de interesse, constitui a metodologia estatística mais adequada à maioria das situações em que se segue uma amostra de indivíduos durante um período de tempo. Esta dissertação pretende abordar a temática dos modelos de regressão para a modelação do tempo de vida quando, para cada indivíduo, se observam acontecimentos múltiplos possivelmente correlacionados.

Neste texto, o capítulo 1 dedicar-se-á à definição de alguns conceitos de análise de sobrevivência. A variável aleatória tempo de vida será definida, assim como a sua caracterização por intermédio das funções de sobrevivência e de risco. Serão apresentados estimadores não paramétricos destas funções e uma família de testes para a comparação da função de sobrevivência entre grupos de indivíduos.

Sendo o modelo de riscos proporcionais de Cox a metodologia de base dos modelos para acontecimentos múltiplos, o capítulo 2 ser-lhe-á inteiramente dedicado. Serão abordadas metodologias para a obtenção de estimadores dos parâmetros, da matriz de covariância e da função de risco subjacente; a análise dos resíduos será também apresentada. A estratificação do modelo de Cox e a inclusão de variáveis dependentes do tempo são úteis na definição dos modelos de acontecimentos múltiplos, tendo-se por isso dedicado o final deste capítulo a estes temas.

O capítulo 3 começará por introduzir alguns conceitos chave e alguma notação usada nos modelos de acontecimentos múltiplos. Posteriormente serão definidos os modelos mais utilizados: modelo de Andersen-Gill (AG), modelo de Wei, Lin e Weissfeld (WLW), modelo de Prentice, Williams e Peterson (PWP) e o modelo de Lee, Wei e Amato (LWA).

A utilização do modelo PWP para a determinação das covariáveis com efeito significativo no tempo até à ocorrência de enfartes múltiplos será apresentada no capítulo 4, com base num estudo real envolvendo doentes hospitalizados entre 1 de janeiro de 2004 e 1 de setembro de 2006 na Unidade de Cuidados Intensivos de Cardiologia dos Hospitais da Universidade de Coimbra.

1
Conceitos básicos de análise de sobrevivência

1.1 Introdução

Os estudos longitudinais prospetivos são geralmente dispendiosos e difíceis de realizar por exigirem uma observação regular dos indivíduos durante um certo período de tempo. Por vezes a duração do estudo é longa e alguns indivíduos acabam por abandoná-lo, por mudança de residência, recusa em continuar o estudo, aparecimento de efeitos adversos provocados por um medicamento ou mesmo o óbito. Outras vezes o estudo não dura o tempo suficiente para se observar o acontecimento para todos os participantes. A análise de sobrevivência veio assim desempenhar um papel fundamental na análise de dados em que pode não ser possível observar, para todos os indivíduos da amostra, o instante em que o acontecimento se realizou e, portanto, o seu tempo de vida.

A designação "análise de sobrevivência" deve-se ao facto desta metodologia estatística ter sido inicialmente desenvolvida com o objetivo de estudar o tempo até à ocorrência do acontecimento morte, daí os tempos serem designados por tempos de sobrevivência ou tempos de vida.

Nas quatro últimas décadas a análise de sobrevivência tem tido uma posição de destaque, quer pelo elevado número de métodos desenvolvidos, quer pelas diversas áreas em que se pode aplicar. Algumas dessas áreas são Psicologia, Demografia, Física, Engenharia ou Medicina. Conforme a área de estudo o tempo de vida pode ser definido como, por exemplo, o tempo até à falha de uma componente mecânica ou elétrica, o tempo até à recaída de um toxicodependente depois de uma desintoxicação ou o tempo até à remissão de um tumor.

Sem perda de generalidade, neste texto serão referidos como tempos de vida todos os tempos até à ocorrência de um acontecimento, assim como a ocorrência de morte significará a ocorrência de um qualquer acontecimento de interesse.

1.2 Censura

Um dos principais motivos pelo qual a análise de sobrevivência é a metodologia mais adequada para analisar o tempo até um determinado acontecimento é a presença de dados censurados. Define-se como uma observação censurada, aquela que corresponde a um indivíduo cujo tempo de vida não foi possível observar, mas em vez disso, se observou o tempo desde a sua entrada no estudo e o instante

do abandono ou fim do estudo. A análise de sobrevivência veio revolucionar a análise destes dados, uma vez que leva em conta a informação relativa ao indivíduo, enquanto ainda é possível observá-lo. De outra forma, a sua exclusão da análise estatística envolveria uma grande perda de informação e produziria resultados enviesados.

Existem vários mecanismos de censura, entre eles a censura à esquerda, censura intervalar e censura à direita. No caso de censura à esquerda, o tempo de vida é inferior ao tempo durante o qual o indivíduo esteve em observação, que é o tempo de censura. A censura intervalar é frequentemente encontrada em estudos em que os indivíduos são seguidos periodicamente e o acontecimento ocorre num intervalo de tempo entre duas entrevistas. Existe censura à direita quando o tempo de vida é superior ao tempo de censura. Neste trabalho será apenas considerada censura à direita.

Do ponto de vista formal, importa introduzir alguma notação. Numa amostra de dimensão n, a cada indivíduo i corresponde o par (T_i^*, δ_i), onde:

- $T_i^* = \min(T_i, C_i)$ designa o tempo de *follow-up*, sendo T_i o tempo de vida (ou o tempo que decorreu desde a entrada em estudo até à realização do acontecimento de interesse) e C_i o tempo de censura (ou o tempo potencial de observação desde a entrada em estudo até à cessação da observação sem que o acontecimento se tenha realizado);

- $\delta_i = \mathrm{I}(\{T_i \leq C_i\})$ denota o estado final do indivíduo, tomando o valor 1 se foi possível observar o tempo de vida T_i e o valor 0 se só foi possível observar o tempo de censura C_i;

Assume-se ainda, de agora em diante, que a censura é não informativa por esta ser independente do mecanismo de falha. Portanto, os indivíduos a que correspondem observações censuradas não foram excluídos do estudo por apresentarem risco de morte excessivamente elevado ou baixo quando comparados com os indivíduos ainda em observação.

1.3 Funções associadas ao tempo de vida

Seja T uma variável aleatória não negativa que representa o tempo de vida de um indivíduo proveniente de uma dada população homogénea. A distribuição de T pode ser univocamente especificada através de qualquer uma das seguintes funções: função de sobrevivência $S(t)$, função densidade de probabilidade $f(t)$ ou função de risco $\lambda(t)$.

Noutras áreas da estatística, uma variável aleatória é habitualmente definida a partir da função densidade de probabilidade ou função massa de probabilidade, e da função de distribuição. Na análise de sobrevivência, as funções fundamentais para o estudo do tempo até à realização do acontecimento de

interesse são a função de sobrevivência e a função de risco.

A definição matemática destas funções será feita em separado para os casos em que a variável tempo é contínua ou discreta.

1.3.1 Caso contínuo

A função a definir em primeiro lugar é a função de distribuição $F(t)$ da variável aleatória T. Esta função representa a probabilidade do tempo de vida ser inferior ou igual a um dado instante t:

$$F(t) = \mathrm{P}(T \leq t),\ 0 \leq t < +\infty.$$

Esta função é monótona não decrescente, contínua à direita e tal que, $F(0) = 0$ e $\lim_{t\to\infty} F(t) = 1$. Quando $F(t)$ é diferenciável, define-se a função densidade de probabilidade de T por

$$F'(t) = f(t) = \lim_{\triangle t \to 0^+} \frac{\mathrm{P}(t < T \leq t + \triangle t)}{\triangle t}.$$

Esta expressão dá um valor aproximado para a probabilidade do acontecimento ocorrer num intervalo infinitesimal $]t; t + \triangle t]$.

A função de sobrevivência $S(t)$ define-se como a probabilidade do indivíduo sobreviver para além do instante t:

$$S(t) = \mathrm{P}(T > t) = \int_t^{+\infty} f(u)\, du,\ 0 \leq t < +\infty.$$

A função de sobrevivência é uma função monótona não crescente, contínua à esquerda e que satisfaz as propriedades: $S(0) = 1$ e $\lim_{t\to+\infty} S(t) = 0$. Além disso, $f(t) = F'(t) = -S'(t)$.

A função de risco representa a taxa instantânea de ocorrência do acontecimento no instante t, sabendo que este não ocorreu até esse instante e define-se por:

$$\lambda(t) = \lim_{\triangle t \to 0^+} \frac{\mathrm{P}(t < T \leq t + \triangle t \mid T > t)}{\triangle t}.$$

A função de risco satisfaz as seguintes propriedades, $\lambda(t) \geq 0$ e $\int_0^{+\infty} \lambda(t)\, dt = +\infty$. O comportamento da função de risco está associado à situação que se está a analisar; pode ser monótona crescente, monótona decrescente, constante ou apresentar padrões mais complexos como no caso do risco de morte em indivíduos tuberculosos. Neste caso, o risco aumenta de forma muito acentuada nos primeiros dias após o diagnóstico e diminui rapidamente após a administração do tratamento.

A função de risco cumulativa ou integrada é definida por

$$\Lambda(t) = \int_0^t \lambda(u)\, du.$$

As três funções anteriores relacionam-se entre si. Uma das relações mais utilizadas é:

$$\lambda(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt} \ln S(t). \tag{1}$$

Dado que $S(0) = 1$, obtém-se facilmente a relação:

$$\Lambda(t) = -\ln S(t) \quad \Longleftrightarrow \quad S(t) = \exp[-\Lambda(t)]. \tag{2}$$

1.3.2 Caso discreto

Quando os tempos de vida estão agrupados ou se referem a um número inteiro de ciclos de determinada natureza, pode ser preferível considerar T uma variável aleatória discreta a tomar os valores, $t_i, i = 1, 2, ...$, com $t_1 < t_2 < \cdots$. No caso discreto, em vez de se falar de função densidade de probabilidade, define-se a função massa de probabilidade que a cada realização t_i de T faz corresponder a probabilidade de ocorrência de um acontecimento nesse instante,

$$f_i = \mathrm{P}\left(T = t_i\right) > 0, \quad i = 1, 2, ...$$

A definição da função de sobrevivência é idêntica à definição apresentada no caso contínuo,

$$S(t) = 1 - F(t) = 1 - \mathrm{P}(T \leq t) = \mathrm{P}(T > t) = \sum_{i:t_i > t} f_i,$$

a partir da qual se obtém a relação, $f_j = S(t_{j-1}) - S(t_j)$. O valor de $S(t)$ apenas varia em cada instante t_j. Imediatamente antes de qualquer instante t_j, ou seja, em t_j^-, o valor de $S(t)$ é superior ao observado em t_j:

$$S(t_j^-) > S\left(t_j\right) \quad \text{e} \quad S(t_j^-) = S(t_{j-1}).$$

Isto significa que a função de sobrevivência é uma função em escada, monótona não crescente e contínua à direita.

A função de risco é uma função que a cada instante t_j faz corresponder um valor λ_j que representa a probabilidade condicional de ocorrência do acontecimento em t_j, $\lambda_j = \mathrm{P}(T = t_j \mid T > t_j^-)$, que, analogamente ao que acontecia no caso contínuo, é tal que

$$\lambda_j = \frac{f_j}{S(t_j^-)} = \frac{S(t_{j-1}) - S(t_j)}{S(t_{j-1})} \iff S(t_j) = S(t_{j-1})(1 - \lambda_j).$$

Aplicando esta fórmula recursivamente para $t_1 < t_2 < \cdots$, e considerando que $S(t_0) = S(0) = 1$, tem-se que

$$S(t) = \prod_{j:t_j \leq t} (1 - \lambda_j).$$

A partir da relação (2) entre $\Lambda(t)$ e $S(t)$ obtém-se

$$\Lambda(t) = -\ln S(t) = -\sum_{j:t_j \leq t} \ln(1-\lambda_j).$$

No entanto, se λ_j for suficientemente pequeno, $\ln(1-\lambda_j)$ é aproximadamente igual a $-\lambda_j$ e surge uma expressão alternativa para $\Lambda(t)$:

$$\Lambda(t) = \sum_{j:t_j \leq t} \lambda_j.$$

1.4 Alguns estimadores não paramétricos

1.4.1 Estimador de Kaplan-Meier

O estimador desenvolvido por Kaplan e Meier (1958) para a função de sobrevivência também faz uso da informação relativa aos indivíduos cuja morte não é observada. Para se sobreviver t unidades de tempo desde um instante inicial t_0, é preciso que se sobreviva a todos os instantes entre t_0 e $t_0 + t$. A probabilidade de sobrevivência a um instante t é dada pela razão entre o número de indivíduos que sobrevivem ao instante t e o número de indivíduos que estão em risco imediatamente antes desse instante, admitindo que os indivíduos apresentam igual probabilidade de morte. Consideram-se em risco os indivíduos censurados em t, os sobreviventes a t e aqueles aos quais ocorre o acontecimento em t.

Considere-se uma amostra de dimensão n proveniente de uma população homogénea, ou seja, em que os indivíduos não diferem quanto ao risco de morte. Nesta amostra, m é o número de tempos de vida distintos tais que, $0 = t_0 < t_1 < \cdots < t_m$, com $m \leq n$. Formalmente, a probabilidade de sobreviver para além do instante t_i, $i = 1, \cdots, m$, pode ser escrita como

$$S(t_i) = \mathrm{P}(T > t_i) = \mathrm{P}(T > t_{i-1}) \times p_i = S(t_{i-1}) \times p_i = \prod_{k=1}^{i} p_k,$$

com $p_i = \mathrm{P}(T > t_i \mid T > t_{i-1}) = \dfrac{\mathrm{P}(T > t_i)}{\mathrm{P}(T > t_{i-1})}$.

Como apenas se estão a considerar instantes em que ocorreram mortes, a cada instante t_i está associado o número de mortes ocorridas, $d_i \geq 1, i = 1, \cdots, m$. Diz-se que as observações são empatadas quando $d_i > 1$. Para cada instante t_i define-se o número n_i de indivíduos em risco em t_i^-. Por convenção, estabeleceu-se que em instantes em que ocorram mortes e censuras, os indivíduos censurados são

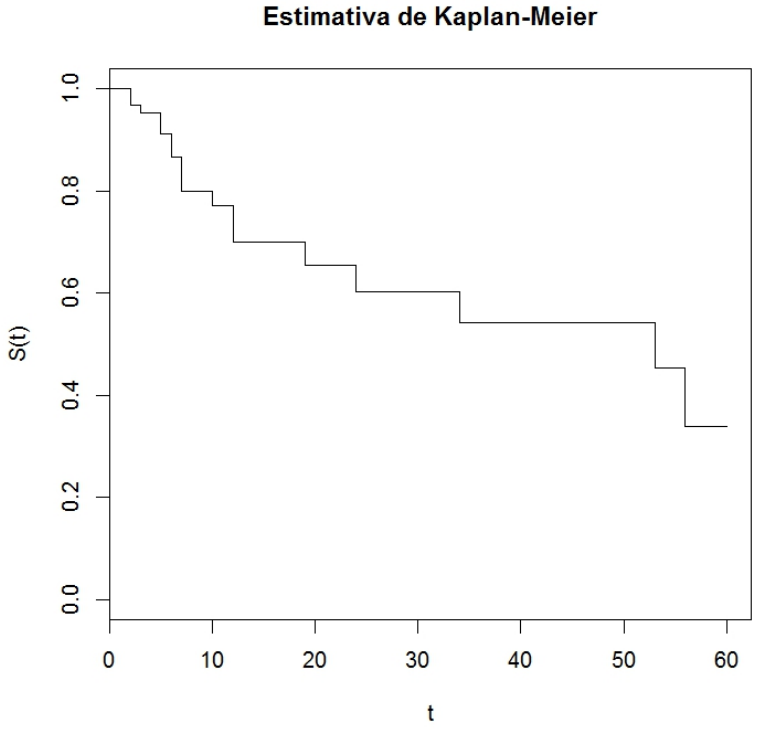
incluídos no grupo de indivíduos em risco nesse instante. O estimador da probabilidade condicional p_i de sobreviver ao instante t_i é dado por,

$$\widehat{p}_i = \frac{n_i - d_i}{n_i} = 1 - \frac{d_i}{n_i} = 1 - \widehat{q}_i, \qquad i = 1, \cdots, m,$$

onde $\widehat{q}_i$ é o estimador da probabilidade de morte em t_i, dada a sobrevivência até esse instante.
O estimador de Kaplan-Meier para a função de sobrevivência no instante t é

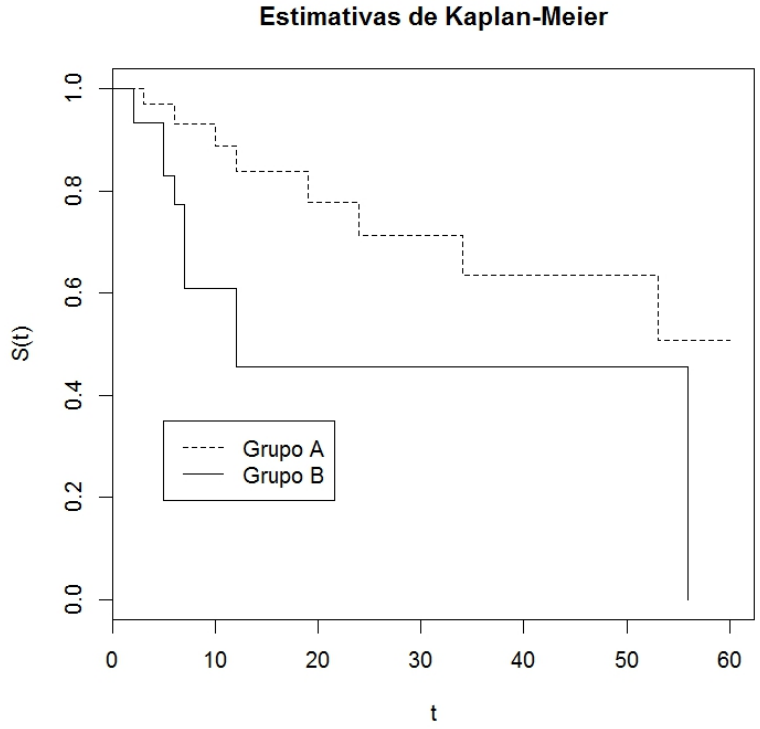
$$\widehat{S}_{KM}(t) = \prod_{j:t_j \leq t} \left(1 - \frac{d_j}{n_j}\right). \tag{3}$$

As estimativas obtidas a partir do estimador de Kaplan-Meier podem ser representadas graficamente. A curva resultante é o gráfico de uma função em escada que decresce em cada instante de morte. A altura dos degraus é tanto maior quanto maior for o número de censuras ocorridas desde a última morte, visto que estas provocam uma redução do número de indivíduos em risco. Seja t^* o último tempo observado; o estimador de Kaplan-Meier toma o valor zero neste instante e para além dele quando t^* é um tempo de vida. No entanto, quando t^* é uma observação censurada, o estimador não toma o valor zero e não está definido para $t > t^*$.
A figura seguinte mostra um exemplo de uma curva de sobrevivência.



Se os indivíduos apresentarem alguma característica que os distinga, a estimativa de Kaplan-Meier deve ser obtida para cada grupo homogéneo de indivíduos e representada num gráfico. Assim, é possível

observar se os indivíduos de um dos grupos apresentam uma maior sobrevivência. A partir da curva que se segue, observa-se que nesta amostra os indivíduos do Grupo B tendem a morrer mais rapidamente do que os indivíduos do Grupo A.



O estimador de Kaplan-Meier não permite estudar o efeito em simultâneo de diversas covariáveis no tempo de vida, nem ajustar esse efeito a eventuais confundimentos de outras variáveis. Nesse caso deve usar-se um modelo de regressão, como por exemplo o modelo de riscos proporcionais de Cox.

1.4.2 Estimador de Nelson-Aalen

Nesta secção vai-se introduzir o estimador de Nelson-Aalen da função de risco cumulativa. Como a função de sobrevivência pode ser estimada pelo estimador de Kaplan-Meier, um estimador natural para a função de risco cumulativa é $\widehat{\Lambda}(t) = -\ln \widehat{S}_{KM}(t)$, mas este não é o mais usual.
Dada uma amostra de dimensão n com m tempos de vida distintos, $t_1 < ... < t_i < \cdots < t_m$, com $m \leq n$, o estimador de Nelson-Aalen ou função de risco cumulativa empírica é,

$$\widehat{\Lambda}_{NA}(t) = \sum_{j:t_j \leq t} \frac{d_j}{n_j}. \tag{4}$$

Para amostras de grande dimensão e com poucos valores empatados, os dois estimadores apresentados acima são assintoticamente equivalentes e não diferem muito na maior parte dos casos. No entanto,

nos instantes de tempo finais, em que existem poucos indivíduos em risco, as estimativas podem diferir de forma mais marcante.

1.4.3 Estimador de Breslow

Breslow (1972) sugeriu que se considerasse o estimador de Nelson-Aalen para a função de risco cumulativa e se usasse a relação (2) para determinar outro estimador não paramétrico da função de sobrevivência

$$\widehat{S}_B(t) = \exp[-\widehat{\Lambda}_{NA}(t)] = \exp\left(-\sum_{j:t_j \leq t} \frac{d_j}{n_j}\right) = \prod_{j:t_j \leq t} \mathrm{e}^{-(d_j/n_j)}, \tag{5}$$

onde d_j e n_j representam, respetivamente, o número de mortes e o número de indivíduos em risco no instante t_j.

Foi provado por Fleming e Harrington (1984) que os estimadores da função de sobrevivência propostos por Kaplan-Meier e Breslow são assintoticamente equivalentes para amostras suficientemente grandes. A comparação entre os dois estimadores foi realizada para diversas dimensões de amostra e diversas percentagens de censura. Tem-se que

$$\widehat{S}_B(t) = \prod_{j:t_j \leq t} \mathrm{e}^{-(d_j/n_j)} \approx \prod_{j:t_j \leq t} \left(1 - \frac{d_j}{n_j}\right) = \widehat{S}_{KM}(t),$$

dado que para valores pequenos de x, $\mathrm{e}^{-x} \approx 1 - x$. Logo, os estimadores são bastante próximos quando d_j/n_j é pequeno, isto é, quando ainda existem muitos indivíduos em risco. De notar que para amostras finitas, $\mathrm{e}^{-x} \geq 1 - x$, e portanto, $\widehat{S}_B(t) \geq \widehat{S}_{KM}(t)$.

Quando o último tempo observado é um tempo de vida, $\widehat{S}_{KM}(t) = 0$ enquanto $\widehat{S}_B(t) > 0$. Em geral, nas caudas das curvas de sobrevivência, obtidas pelos dois estimadores, as diferenças são mais acentuadas por existirem poucos indivíduos em risco, mas este é um problema inerente à própria estimação não paramétrica da função de sobrevivência.

Os estimadores podem também apresentar diferenças consideráveis na presença de muitas observações empatadas. Neste caso, o melhor será optar pela estimativa de Kaplan-Meier.

1.5 Família de testes Tarone-Ware

Além da determinação de estimativas para as diferentes funções da análise de sobrevivência, é também habitual a sua comparação entre dois ou mais grupos de indivíduos formados de acordo com os valores que uma característica pode tomar. No caso em que indivíduos são distribuídos por dois grupos, A e

B, o mais frequente é testar as hipóteses sobre as respetivas funções de sobrevivência:

$$H_0 : S_A(t) = S_B(t) \qquad vs \qquad H_a : S_A(t) \neq S_B(t).$$

Uma vez que a função de distribuição de T é $1-S(t)$, quando os dados não são censurados, as hipóteses anteriores podem ser testadas usando testes não paramétricos, como por exemplo o teste de Mann-Whitney-Wilcoxon. Dada a existência de dados censurados no âmbito da análise de sobrevivência a utilização dos testes usuais fica comprometida, o que motivou o desenvolvimento de uma família de testes como uma extensão dos já existentes, mas que permitisse a inclusão de dados censurados.
Considere-se uma amostra de dimensão $n = n_A + n_B$, em que n_A e n_B representam, respetivamente, as dimensões dos grupos A e B. Sejam $t_1 < t_2 < \cdots < t_J$ os J instantes de morte observados nos n indivíduos. Em cada instante $t_j, j = 1 \cdots, J$ os dados podem organizar-se segundo a tabela de contingência

Grupo	Número de mortes em t_j	Número de sobreviventes em t_j	Número de indivíduos em risco em t_j^-
A	d_{Aj}	$n_{Aj} - d_{Aj}$	n_{Aj}
B	d_{Bj}	$n_{Bj} - d_{Bj}$	n_{Bj}
Total	d_j	$n_j - d_j$	n_j

Assume-se que $d_{A1}, d_{A2}, \cdots, d_{AJ}$ são variáveis aleatórias independentes tais que $\mathrm{E}(d_{Aj}) = n_{Aj}(d_j/n_j)$ e $\mathrm{var}(d_{Aj}) = [n_{Aj}n_{Bj}d_j(n_j - d_j)]/[n_j^2(n_j - 1)]$. Dadas as constantes conhecidas, $w_1, w_2, \cdots, w_J$, a estatística de teste

$$Q = \frac{\left\{\sum_{j=1}^{J} w_j \left[d_{Aj} - \mathrm{E}(d_{Aj})\right]\right\}^2}{\sum_{j=1}^{J} w_j^2 \ \mathrm{var}(d_{Aj})}$$

tem distribuição assintótica χ_1^2 sob a hipótese nula de igualdade das funções de sobrevivência.
Existe uma diversidade de estatísticas de teste que diferem conforme a escolha das constantes w_j. As escolhas mais comuns são:

$w_j = 1$	Teste Log-rank,
$w_j = n_j$	Teste de Gehan,
$w_j = \sqrt{n_j}$	Teste de Tarone e Ware,
$w_j = \prod_{l=1}^{j} \frac{n_l - d_l + 1}{n_l + 1}$	Teste Peto-Prentice.

A estatística do teste Log-rank atribui um peso constante às J diferenças entre o número observado e o número esperado de mortes, enquanto a estatística do teste de Gehan dá maior peso às diferenças observadas nos primeiros instantes de morte. Tarone e Ware (1977), por seu lado, argumentaram que os pesos definidos por Gehan (1965) perdiam sensibilidade quando as distribuições dos tempos censurados diferiam significativamente entre os dois grupos de indivíduos e sugeriram uma definição que fosse

um compromisso entre os pesos dos testes Log-rank e de Gehan. A estatística do teste Peto-Prentice surge também como uma tentativa de solucionar as limitações observadas nas estatísticas dos testes Log-rank e de Gehan, considerando pesos que são valores muito próximos da estimativa de Kaplan-Meier da função de sobrevivência comum aos dois grupos de indivíduos.
Note-se que o teste Log-rank é o teste mais potente quando as funções de risco são proporcionais e não deve ser usado quando existe cruzamento das funções de risco.

Quando existem mais de dois grupos, as hipóteses

$$H_0 : S_1(t) = S_2(t) = \cdots = S_g(t) = \cdots = S_G(t) \qquad vs \qquad H_a : \exists g_1, g_2 \text{ tal que } S_{g_1}(t) \neq S_{g_2}(t).$$

podem ser testadas usando uma extensão das estatísticas de teste consideradas no caso em que se têm dois grupos. Assim, para um instante de morte $t_j, j = 1 \cdots, J$ e um grupo $g = 1 \cdots, G$, define-se d_{gj} como o número de mortes ocorridas em t_j, e n_{gj} o número de indivíduos em risco imediatamente antes de t_j. Seja $\mathrm{E}(d_{gj}) = n_{gj}d_j/n_j$ o número de mortes esperadas no grupo g no instante t_j, $\mathbf{U} = (U_1, \cdots, U_g, \cdots, U_{G-1})$ é tal que

$$U_g = \sum_{j=1}^{J} w_j[d_{gj} - \mathrm{E}(d_{gj})], \quad g = 1, \cdots, G-1.$$

A matriz de covariância $\mathbf{V}$ de dimensão $(G-1)\times(G-1)$ é uma matriz simétrica, cujo elemento genérico (g, l) é

$$\mathbf{V}_{g,l} = \begin{cases} \sum_{j=1}^{J} w_j^2 \frac{n_{gj}(n_j - n_{gj})d_j(n_j - d_j)}{n_j^2(n_j - 1)}, g = l \\ \sum_{j=1}^{J} -w_j^2 \frac{n_{gj}n_{lj}d_j(n_j - d_j)}{n_j^2(n_j - 1)}, \quad g \neq l \end{cases} \qquad g, l = 1, \cdots, G-1.$$

A estatística de teste é $\mathbf{Q} = \mathbf{U}'\mathbf{V}^{-1}\mathbf{U}$ que, sob H_0, tem distribuição assintótica χ^2_{G-1}.

2
Modelo de regressão de Cox

2.1 Introdução

A análise de regressão de Cox é aplicada a estudos longitudinais em que se pretende fazer uma avaliação do prognóstico dos indivíduos em estudo considerando, neste modelo, o tempo de vida como variável resposta e diversos fatores que se pensa afetarem o tempo de vida, como variáveis explanatórias. As variáveis explanatórias podem ser de vários tipos: numéricas contínuas, numéricas discretas ou categóricas.

O modelo de Cox tem grande utilidade quando se pretende quantificar ou estimar o efeito de um tratamento no tempo de vida. Este objetivo pode ser conseguido pela determinação do risco relativo, cuja estimativa é ajustada à presença de eventuais efeitos de confundimento de outras variáveis. A existência de efeitos de confundimento é bastante comum em estudos realizados em medicina, quer sejam eles observacionais ou experimentais.

São consideradas variáveis com efeito de confundimento todas aquelas que mascaram ou interferem no efeito que um fator tratamento ou um fator de risco têm na variável resposta. Geralmente estas variáveis caracterizam os indivíduos como é o caso do género, da idade ou do historial clínico. Na prática, para uma variável ter um efeito de confundimento tem de estar associada simultaneamente à variável resposta e à variável explanatória cujo efeito se pretende testar.

2.2 Definição do modelo de Cox

Qualquer análise de regressão permite modelar a relação existente entre uma variável resposta e variáveis explanatórias, também designadas por covariáveis ou preditores. Na análise de sobrevivência, dada a importância da função de risco, uma opção consiste em definir modelos de regressão baseados nesta função. Assim, dado um vetor $\mathbf{z}$ de covariáveis e $\boldsymbol{\beta}$ um vetor de parâmetros, é frequente considerar-se o modelo de regressão da forma $\lambda(t;\mathbf{z};\boldsymbol{\beta}) = \lambda_0(t) \times r(\mathbf{z};\boldsymbol{\beta})$, tendo Cox (1972) proposto um modelo com $r(\mathbf{z};\boldsymbol{\beta}) = \exp(\boldsymbol{\beta}'\mathbf{z})$.

Considere-se uma amostra de dimensão n em que para cada indivíduo i se observou o vetor $(t_i, \delta_i, \mathbf{z}_i)$, onde t_i representa o tempo de vida, δ_i indica a ocorrência ou não do acontecimento e $\mathbf{z}_i' = (z_{i1}, \cdots, z_{ip})$ representa os valores observados das p covariáveis fixas. Formalmente o modelo de Cox define-se

assumindo que a função de risco no instante t para o indivíduo i com vetor de covariáveis $\mathbf{z}_i$ é

$$\lambda(t;\mathbf{z}_i) = \lambda_0(t)\mathrm{e}^{(\beta_1 z_{i1}+\cdots+\beta_p z_{ip})} = \lambda_0(t)\mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_i} \qquad i = 1,\cdots,n, \tag{6}$$

em que $\boldsymbol{\beta}' = (\beta_1,\cdots,\beta_p)$ é um vetor de parâmetros desconhecidos. Para simplificar a notação escreve-se $\lambda(t;\mathbf{z}_i)$ em vez de $\lambda(t;\mathbf{z}_i;\boldsymbol{\beta})$. Assim, a função de risco do modelo de Cox é constituída por:

1. $\lambda_0(t)$, é uma função não negativa que caracteriza a forma como o tempo modifica o risco de morte de cada indivíduo. Trata-se de uma função não especificada que apenas depende do tempo e que se assume ser comum a todos os indivíduos;
2. $\mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}}$, que representa a influência que as covariáveis têm na função de risco e designa-se por risco relativo. Esta função depende dos valores das p covariáveis observadas nos indivíduos e dos parâmetros $\boldsymbol{\beta}$ que é necessário estimar. É de notar que, quando $\mathbf{z} = \mathbf{0}$, $\mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}} = 1$ e $\lambda(t;\mathbf{z})$ é exatamente igual a $\lambda_0(t)$. Portanto, $\lambda_0(t)$ representa a função de risco para um indivíduo padrão a que corresponde o vetor $\mathbf{z} = \mathbf{0}$, daí que $\lambda_0(t)$ seja designada por função *hazard baseline* ou função de risco subjacente.

Por (6) pode também obter-se a função de sobrevivência para o indivíduo i com vetor de covariáveis $\mathbf{z_i}$ que é dada por

$$S(t;\mathbf{z_i}) = [S_0(t)]^{\exp(\boldsymbol{\beta}'\mathbf{z_i})}. \tag{7}$$

A partir deste modelo é possível definir a razão dos riscos também conhecida por *hazard ratio* ou risco relativo. Dados dois indivíduos j e k com vetores de covariáveis $\mathbf{z}_j$ e $\mathbf{z}_k$, o risco relativo é dado por,

$$\frac{\lambda(t;\mathbf{z}_j)}{\lambda(t;\mathbf{z}_k)} = \frac{\lambda_0(t)\mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_j}}{\lambda_0(t)\mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_k}} = \mathrm{e}^{\boldsymbol{\beta}'(\mathbf{z}_j-\mathbf{z}_k)}.$$

Como a razão anterior não depende do tempo, pode dizer-se que o modelo de Cox é um modelo de riscos proporcionais. Trata-se de um modelo semi-paramétrico porque, apesar de não se especificar a distribuição de $\lambda_0(t)$, este depende do vetor de parâmetros $\boldsymbol{\beta}$.

O vetor de parâmetros é bastante importante, uma vez que β_i representa o efeito da covariável z_i no tempo de vida. Sejam z_{ji} e z_{ki} os valores da covariável z_i correspondentes aos indivíduos j e k, respetivamente. Admite-se que os dois indivíduos considerados diferem apenas em relação à covariável z_i e que $z_{ji} - z_{ki} = 1$. Então o risco de morte do indivíduo j relativamente ao indivíduo k é dado por $RR = \mathrm{e}^{\beta_i}$ e o seu significado depende do seu valor:

RR=1, a covariável não tem um efeito significativo no tempo de vida;

RR>1, a covariável é um preditor de mau prognóstico, o indivíduo j vê o seu risco de morte aumentado em relação ao do indivíduo k. O aumento do risco é dado por $100(RR-1)\%$;

RR<1, a covariável é um preditor de bom prognóstico, o indivíduo j vê o seu risco de morte reduzido em relação ao do indivíduo k. A diminuição do risco é dada por $100(1-RR)\%$.

2.3 Estimação dos parâmetros do modelo

Depois de introduzido o modelo semi-paramétrico de Cox interessa agora estimar o vetor de parâmetros desconhecidos $\boldsymbol{\beta}$. Para isso é necessário construir a função de verosimilhança.

2.3.1 Função de verosimilhança parcial

No caso de se assumir um modelo paramétrico para o tempo de vida, cuja distribuição depende do vetor de parâmetros θ, a função de verosimilhança é o produto de duas componentes: uma referente aos indivíduos cujos tempos de vida foram observados e outra referente aos indivíduos cujos tempos foram censurados. Nesta última situação, os indivíduos contribuem para a verosimilhança através da probabilidade de sobreviverem para além do instante em que saem do estudo. Para uma amostra de dimensão n, a função de verosimilhança é então dada por $L(\boldsymbol{\theta}) = \prod_{i=1}^{n}[f(t_i;\boldsymbol{\theta})]^{\delta_i}[S(t_i;\boldsymbol{\theta})]^{1-\delta_i}$.
No entanto, o modelo de Cox não assume que o tempo de vida apresente uma distribuição paramétrica, e portanto $f(z)$ é desconhecida, o que torna inapropriada a função de verosimilhança anterior. Para solucionar este problema, Cox propôs no seu artigo de 1972 uma função de "verosimilhança" não dependente de $\lambda_0(t)$ e que mais tarde Cox (1975) viria a mostrar ser uma verosimilhança parcial.

Sejam m tempos de vida distintos, $t_1 < t_2 < \cdots < t_m$, observados numa amostra de dimensão n, tal que $n \geq m$. A cada instante t_j corresponde um conjunto de indivíduos em risco, $R(t_j) = R_j$. Dado o vetor de covariáveis $\mathbf{z}_i$ associado ao indivíduo i $(i = 1, \cdots, n)$, a função de verosimilhança parcial é dada por,

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{m} \frac{\mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_j}}{\sum\limits_{i \in R_j} \mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_i}}. \tag{8}$$

Apesar da função de verosimilhança parcial (8) utilizar apenas uma parte dos dados, Cox (1972) argumentou que esta podia ser usada como uma função de verosimilhança, o que permitiu a utilização dos procedimentos habituais para a estimação dos parâmetros. Além disso, foi provado usando a teoria dos processos de contagem, que o estimador de $\boldsymbol{\beta}$ assim obtido goza das propriedades assintóticas usuais dos estimadores de máxima verosimilhança.

2.3.2 Estimador de máxima verosimilhança parcial

O estimador de máxima verosimilhança parcial de $\boldsymbol{\beta}$ é dado pelo $\hat{\boldsymbol{\beta}}$ que maximiza (8). Fica claro que o estimador de $\boldsymbol{\beta}$ não depende dos tempos de vida. Além disso (8) constitui um produto em que apenas se multiplicam os fatores referentes a indivíduos cujo tempo de vida foi observado, os indivíduos com

tempos censurados contribuem para a análise estatística, na medida em que são incluídos, em cada instante t_j, no conjunto em risco R_j. Um indivíduo com tempo censurado igual a um tempo de vida t_k, é ainda considerado em risco nesse instante, ou seja, pertence a R_k, saindo posteriormente do estudo. De forma a simplificar os cálculos, e uma vez que a função logaritmo é monótona crescente, em vez de se determinar o máximo da função definida em (8) determina-se o máximo da função log-verosimilhança parcial,

$$l(\boldsymbol{\beta}) = \ln(L(\boldsymbol{\beta})) = \sum_{j=1}^{m} \left\{ \boldsymbol{\beta}'\mathbf{z}_j - \ln \left[\sum_{i \in R_j} \mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_i} \right] \right\}.$$

O estimador de máxima verosimilhança parcial de $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}' = (\hat{\beta}_1, \cdots, \hat{\beta}_p)$, é obtido a partir de um sistema de p equações. As equações são obtidas derivando a função log-verosimilhança parcial em ordem a cada um dos parâmetros desconhecidos, $\beta_k,\ k = 1, \cdots, p$, e igualando a zero. Vem,

$$\mu_k(\boldsymbol{\beta}) = \frac{\partial l(\boldsymbol{\beta})}{\partial \beta_k} = \sum_{j=1}^{m} \left\{ z_{jk} - \frac{\partial}{\partial \beta_k} \ln \left[\sum_{i \in R_j} \mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_i} \right] \right\} = \sum_{j=1}^{m} \left[z_{jk} - \frac{\sum\limits_{i \in R_j} z_{ik} \mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_i}}{\sum\limits_{i \in R_j} \mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_i}} \right]. \tag{9}$$

O vetor $\boldsymbol{\mu}'(\boldsymbol{\beta}) = (\mu_1(\boldsymbol{\beta}), \cdots, \mu_p(\boldsymbol{\beta}))$ designa-se por vetor *score* e o estimador de máxima verosimilhança parcial de $\boldsymbol{\beta}$ é a solução da equação $\boldsymbol{\mu}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$, o que é equivalente ao seguinte sistema de equações:

$$\begin{cases} \mu_1(\boldsymbol{\beta}) = \frac{\partial l(\boldsymbol{\beta})}{\partial \beta_1} = 0 \\ \cdots \qquad \cdots \qquad \cdots \\ \mu_p(\boldsymbol{\beta}) = \frac{\partial l(\boldsymbol{\beta})}{\partial \beta_p} = 0. \end{cases}$$

A parcela j em (9) representa o desvio do valor z_{jk} da covariável z_k observado no indivíduo j, em relação à média ponderada dos valores de z_k observados em todos os indivíduos em risco no instante t_j, com pesos dados por $\exp(\boldsymbol{\beta}'\mathbf{z}_i)$.

2.4 A matriz de covariância

Antes de se apresentarem as distribuições assintóticas necessárias para a realização de inferência sobre os parâmetros do modelo é importante definir a matriz de covariância.

A matriz de covariância é aproximada a partir da matriz de informação observada,

$$\mathbf{I}(\boldsymbol{\beta}) = -\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2}. \tag{10}$$

O elemento que se encontra na linha r e na coluna k é obtido por

$$-\frac{\partial^2 l(\boldsymbol{\beta})}{\partial\beta_r\partial\beta_k} = -\sum_{j=1}^{m}\frac{\partial}{\partial\beta_r}\left[z_{jk} - \frac{\sum\limits_{i\in R_j} z_{ik}\mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_i}}{\sum\limits_{i\in R_j}\mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_i}}\right]. \tag{11}$$

Quando $r = k$, o k-ésimo elemento da diagonal da matriz é

$$-\frac{\partial^2 l(\boldsymbol{\beta})}{\partial\beta_k^2} = \sum_{j=1}^{m}\left[\frac{\sum\limits_{i\in R_j} z_{ik}^2\mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_i}}{\sum\limits_{i\in R_j}\mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_i}} - \left(\frac{\sum\limits_{i\in R_j} z_{ik}\mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_i}}{\sum\limits_{i\in R_j}\mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_i}}\right)^2\right].$$

Quando $r \neq k$, o elemento (r,k) ou (k,r) da matriz é

$$-\frac{\partial^2 l(\boldsymbol{\beta})}{\partial\beta_r\partial\beta_k} = \sum_{j=1}^{m}\left[\frac{\sum\limits_{i\in R_j} z_{ik}z_{ir}\mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_i}}{\sum\limits_{i\in R_j}\mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_i}} - \frac{\sum\limits_{i\in R_j} z_{ik}\mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_i}\sum\limits_{i\in R_j} z_{ir}\mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_i}}{\left(\sum\limits_{i\in R_j}\mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_i}\right)^2}\right].$$

A matriz de covariância do estimador de $\boldsymbol{\beta}$ é aproximada por,

$$\mathrm{var}(\hat{\boldsymbol{\beta}}) \cong \mathbf{I}^{-1}(\hat{\boldsymbol{\beta}}). \tag{12}$$

O k-ésimo elemento da diagonal da matriz $\mathbf{I}^{-1}(\hat{\boldsymbol{\beta}})$, $\hat{\sigma}_{kk}^2$, corresponde ao estimador da variância de $\hat{\beta}_k$. De notar que nesta matriz os elementos (k,r) ou (r,k) representam o estimador da covariância entre quaisquer dois estimadores $\hat{\beta}_k$ e $\hat{\beta}_r$ para $k \neq r$, o que permite estimar a sua correlação. $\hat{\sigma}_{kk}$ é o estimador do erro padrão do estimador de máxima verosimilhança parcial, $\widehat{\mathrm{EP}}(\hat{\beta}_k)$.

2.5 Inferência sobre os parâmetros

Para fazer inferência sobre os parâmetros de regressão, tanto pela obtenção de intervalos de confiança como pela realização de testes de hipóteses, é necessário conhecer a distribuição de amostragem do estimador de máxima verosimilhança parcial $\hat{\boldsymbol{\beta}}$. Nem sempre é possível obter a distribuição de amostragem exata para este estimador e por isso tem de se recorrer à teoria assintótica, pressupondo a verificação de certas condições de regularidade. Ficou provado matematicamente, recorrendo aos processos de contagem baseados na teoria das martingalas, que os estimadores obtidos a partir da função de verosimilhança parcial apresentam as mesmas propriedades distribucionais dos estimadores obtidos pela função de verosimilhança completa [Anderson, Borgan, Gill e Keiding (1993, Capítulo VII), e

Fleming e Harrington (1991, capítulo 4)].

Pode então dizer-se que:

1. $\hat{\boldsymbol{\beta}}$ tem distribuição assintótica normal p-variada, $\hat{\boldsymbol{\beta}} \overset{\text{a}}{\sim} N_p(\boldsymbol{\beta}, \mathbf{I}^{-1}(\boldsymbol{\beta}))$;
2. O estimador $\hat{\boldsymbol{\beta}}$ é assintoticamente centrado já que $\mathrm{E}(\hat{\boldsymbol{\beta}}) \approx \boldsymbol{\beta}$;
3. Sob a hipótese $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, a estatística de Wald satisfaz, $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)' \, \mathbf{I}(\boldsymbol{\beta}_0) \, (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \overset{\text{a}}{\sim} \chi^2_p$.

Assim, conclui-se que $\hat{\beta}_k \overset{\text{a}}{\sim} N(\beta_k, \mathrm{EP}(\hat{\beta}_k))$.

2.5.1 Testes de hipóteses

Considerando novamente um modelo com p covariáveis e portanto p parâmetros de regressão, é possível realizar testes de hipóteses sobre estes com o objetivo de testar se as covariáveis têm efeito significativo no tempo de vida. Os mais comuns são:

1. Teste da hipótese do efeito nulo da covariável z_j, no tempo de vida,

$$H_0 : \beta_j = 0 \qquad vs \qquad H_a : \beta_j \neq 0. \tag{13}$$

2. Teste da hipótese do efeito nulo de um subconjunto de covariáveis de dimensão $r \leq p$, no tempo de vida,

$$H_0 : \boldsymbol{\beta}_r = \mathbf{0} \qquad vs \qquad H_a : \boldsymbol{\beta}_r \neq \mathbf{0}. \tag{14}$$

Estes testes permitem comparar modelos aninhados, ou seja, comparar submodelos do modelo original. Em (13) compara-se um modelo com p covariáveis com um modelo com $p-1$ covariáveis, no qual foi removida a covariável z_j cujo parâmetro correspondente é β_j. Em (14) compara-se o modelo original com um submodelo com $p-r$ covariáveis, com $1 \leq r \leq p$. Quando $r = p$, testa-se a hipótese de que nenhuma das p covariáveis tem um efeito significativo no tempo de vida.

A comparação de modelos aninhados está na base da seleção de variáveis a incluir no modelo de regressão final. Esta temática será abordada mais tarde.

Os testes mais usuais para testar as hipóteses anteriores são: o teste de Wald, o teste razão de verosimilhanças e o teste *score*.

Teste de Wald

Num modelo com p covariáveis, a estatística do teste de Wald a aplicar quando se pretende testar as hipóteses em (13) é dada por,

$$\mathrm{W} = \frac{\hat{\beta}_j^2}{\widehat{\mathrm{var}}(\hat{\beta}_j)}. \tag{15}$$

Como, sob H_0, $\hat{\beta}_j \overset{a}{\sim} N(0, \text{EP}(\hat{\beta}_j))$, então $\text{W} \overset{a}{\sim} \chi^2_1$.

Para comparar modelos aninhados em que um inclui p covariáveis e o outro inclui apenas $p-r$ covariáveis, a que correspondem as hipóteses em (14), a estatística de teste de Wald é,

$$\text{W} = \hat{\boldsymbol{\beta}}' \, \mathbf{I}(\hat{\boldsymbol{\beta}}) \, \hat{\boldsymbol{\beta}}, \tag{16}$$

onde $\hat{\boldsymbol{\beta}}$ é um vetor de dimensão r e $\mathbf{I}(\hat{\boldsymbol{\beta}})$ é a submatriz simétrica de dimensão $r \times r$ referente aos parâmetros considerados em H_0. De realçar que (13) é um caso particular de (14) quando $r = 1$.

Sob $H_0 : \boldsymbol{\beta}_r = \mathbf{0}$, $\text{W} \overset{a}{\sim} \chi^2_r$, com $1 \leq r \leq p$. A hipótese nula é rejeitada ao nível de significância α se o valor observado de W for superior ao quantil $1-\alpha$ de χ^2_r.

Teste razão de verosimilhanças

Para aplicar o teste razão de verosimilhanças, é necessário determinar os estimadores de máxima verosimilhança parcial sob H_0 e sob H_a. No primeiro caso, o estimador é um vetor com dimensão $p-r$ e representa-se por $\tilde{\boldsymbol{\beta}}$. No segundo caso, o estimador é um vetor de p componentes e representa-se por $\hat{\boldsymbol{\beta}}$. A estatística de teste é dada por

$$\text{G} = -2 \left\{ l(\tilde{\boldsymbol{\beta}}) - l(\hat{\boldsymbol{\beta}}) \right\}. \tag{17}$$

Sob H_0, a distribuição assintótica desta estatística é um χ^2 com número de graus de liberdade igual à diferença entre o número de parâmetros dos modelos considerados em H_a e H_0.

Assim, a hipótese nula é rejeitada ao nível de significância α, se o valor observado de G for superior ao quantil $1-\alpha$ de χ^2_r, em que $r = p - (p-r)$.

Teste *score*

Seja $\hat{\boldsymbol{\beta}}$ o estimador de máxima verosimilhança parcial sob H_a. Seja $\tilde{\boldsymbol{\beta}}$ um vetor que toma o valor zero para os r parâmetros considerados em H_0 e que para os restantes parâmetros é tal que $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$. A estatística do teste *score* é,

$$\text{U} = \boldsymbol{\mu}(\tilde{\boldsymbol{\beta}})' \mathbf{I}^{-1}(\tilde{\boldsymbol{\beta}}) \boldsymbol{\mu}(\tilde{\boldsymbol{\beta}}), \tag{18}$$

onde $\boldsymbol{\mu}(\boldsymbol{\beta})$ é o estimador da função *score* definida em (9). Também aqui a hipótese nula é rejeitada ao nível de significância α, se o valor observado de U for superior ao quantil $1-\alpha$ de χ^2_r.

É habitual usar-se o teste de Wald para testar individualmente se cada parâmetro é zero. Este procedimento torna-se bastante útil quando se está a construir o modelo, porque permite escolher as possíveis variáveis a serem eliminadas do modelo;

Os três testes costumam produzir resultados muito similares; no entanto, quando isso não acontece, deverá optar-se pelo teste de razão de verosimilhanças para tirar conclusões.

2.5.2 Intervalos de confiança

O intervalo de confiança assintótico para cada um dos p parâmetros desconhecidos é obtido tendo em conta a distribuição assintótica de $\hat{\boldsymbol{\beta}}$. O intervalo de $100(1-\alpha)\%$ de confiança para β_k $(k = 1, \cdots, p)$ é dado por

$$\left(\hat{\beta}_k - z_{1-\alpha/2}\widehat{\mathrm{EP}}(\hat{\beta}_k)\ ;\ \hat{\beta}_k + z_{1-\alpha/2}\widehat{\mathrm{EP}}(\hat{\beta}_k)\right), \tag{19}$$

onde $\widehat{\mathrm{se}}(\hat{\beta}_k)$ é a estimativa do erro padrão de $\hat{\beta}_k$ e $z_{1-\alpha/2}$ é o quantil $1-\alpha/2$ da distribuição $N(0,1)$. Por vezes existe interesse em determinar o intervalo de confiança para $c\beta_k, c \in \Re$. Como a distribuição assintótica de $c\hat{\beta}_k$ é $N(c\beta_k, |c|\sigma_{kk})$, o intervalo de confiança para $c\beta_k$ $(k = 1, \cdots, p)$ é

$$\left(c\hat{\beta}_k - z_{1-\alpha/2}|c|\widehat{\mathrm{EP}}(\hat{\beta}_k)\ ;\ c\hat{\beta}_k + z_{1-\alpha/2}|c|\widehat{\mathrm{EP}}(\hat{\beta}_k)\right). \tag{20}$$

A partir de (19) pode obter-se o intervalo de $100(1-\alpha)\%$ de confiança para o risco relativo de morte para dois indivíduos i e j com vetores de covariáveis, $\mathbf{z}_i = (z_{i1}, \cdots, z_{ik}, \cdots, z_{ip})'$ e $\mathbf{z}_j = (z_{j1}, \cdots, z_{jk}, \cdots, z_{jp})'$, que apenas diferem em relação à covariável z_k numa unidade,

$$\left(\mathrm{e}^{\left(\hat{\beta}_k - z_{1-\alpha/2}\widehat{\mathrm{EP}}(\hat{\beta}_k)\right)}\ ;\ \mathrm{e}^{\left(\hat{\beta}_k + z_{1-\alpha/2}\widehat{\mathrm{EP}}(\hat{\beta}_k)\right)}\right).$$

2.6 Estimação de $\lambda_0(t)$, $\Lambda_0(t)$ e $S_0(t)$

Após a estimação dos parâmetros de regressão, interessa por vezes estimar a função de risco subjacente não especificada, $\lambda_0(t)$.

Dados os tempos de vida $t_1 < \cdots < t_m$ com instante inicial de observação $t_0 = 0$, assuma-se que a distribuição do tempo de vida tem uma função de risco constante entre quaisquer dois instantes de morte consecutivos, t_{j-1} e t_j, e que entre eles todos os tempos censurados se consideram ocorridos em t_{j-1}. Assim, a estimativa de $\lambda_0(t)$ no intervalo $(t_{j-1}, t_j]$ é dada por

$$\hat{\lambda_j} = \frac{d_j}{(t_j - t_{j-1}) \sum\limits_{i \in R_j} \mathrm{e}^{\hat{\boldsymbol{\beta}}' \mathbf{z}_i}}, \tag{21}$$

em que d_j e R_j representam, respetivamente, o número de mortes ocorridas e o conjunto de indivíduos em risco no instante t_j. O valor de $\hat{\lambda_j}$ representa o quociente entre o número de acontecimentos ocorridos e o número ponderado de pessoas em risco por unidade de tempo. Cada indivíduo em R_j contribui com um peso $\exp(\hat{\boldsymbol{\beta}}'\mathbf{z}_i)$ no intervalo de tempo considerado.

O estimador de Breslow para a função de risco cumulativa subjacente em cada instante t é dado por,

$$\hat{\Lambda}_0(t) = \sum_{t_j \leq t} \frac{d_j}{\sum\limits_{i \in R_j} \mathrm{e}^{\hat{\boldsymbol{\beta}}'\mathbf{z}_i}}. \tag{22}$$

Pela relação (2), em cada instante t, obtém-se então o estimador da função de sobrevivência subjacente,

$$\hat{S}_0(t) = \exp(-\hat{\Lambda}_0(t)) = \prod_{t_j \leq t} \exp\left(-\frac{d_j}{\sum\limits_{i \in R_j} \mathrm{e}^{\hat{\boldsymbol{\beta}}'\mathbf{z}_i}}\right). \tag{23}$$

Tendo-se obtido o estimador anterior, pode-se então estimar a função de sobrevivência para um indivíduo com vetor de covariáveis $\mathbf{z}$ a partir de (7),

$$\hat{S}(t;\mathbf{z}) = \left[\hat{S}_0(t)\right]^{\exp(\hat{\boldsymbol{\beta}}'\mathbf{z})}.$$

2.7 Observações empatadas

A função de verosimilhança parcial apresentada em (8) foi considerada sob a condição dos tempos de vida observados serem todos distintos. A unidade de tempo escolhida é uma das razões para o pouco rigor na determinação do tempo e consequentemente para o aparecimento de tempos empatados. Quando há observações empatadas é necessário modificar a função de verosimilhança parcial, Kalbfleisch e Prentice (1980) propuseram uma função de verosimilhança para este caso, mas esta é muito exigente do ponto de vista computacional. Felizmente existem aproximações desta função que exigem menor esforço computacional e que foram propostas por, Cox (1972), Peto (1972), Breslow (1974) e Efron (1977).

Sejam m tempos de vida distintos, $t_1 < t_2 < \cdots < t_m$, observados numa amostra de dimensão $n \geq m$, onde a cada indivíduo i $(i = 1, \cdots, n)$ corresponde um vetor de covariáveis $\mathbf{z}_i$. Sejam d_j o número de mortes no instante t_j, $\mathbf{a}_j$ a soma dos vetores de covariáveis correspondentes aos d_j indivíduos e R_j o conjunto de indivíduos em risco nesse instante. A aproximação proposta por Breslow para a função de

verosimilhança parcial é,

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{m} \frac{\mathrm{e}^{\boldsymbol{\beta}'\mathbf{a}_j}}{\left[\sum_{i \in R_j} \mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_i}\right]^{d_j}}. \tag{24}$$

Com função log-verosimilhança parcial,

$$l(\boldsymbol{\beta}) = \sum_{j=1}^{m} \left\{ \boldsymbol{\beta}'\mathbf{a}_j - d_j \ln \left[\sum_{i \in R_j} \mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_i}\right] \right\} \tag{25}$$

e derivadas parciais em ordem a β_k,

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_k} = \sum_{j=1}^{m} \left[a_{jk} - d_j \frac{\sum_{i \in R_j} z_{ik} \mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_i}}{\sum_{i \in R_j} \mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_i}} \right]$$

$$\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_k^2} = -\sum_{j=1}^{m} d_j \left[\frac{\sum_{i \in R_j} z_{ik}^2 \mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_i}}{\sum_{i \in R_j} \mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_i}} - \left(\frac{\sum_{i \in R_j} z_{ik} \mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_i}}{\sum_{i \in R_j} \mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_i}} \right)^2 \right]$$

A partir das expressões anteriores é possível ajustar o modelo de Cox e fazer inferência sobre os seus parâmetros de forma análoga ao que foi exposto nas secções anteriores. Quando $d_j = 1$, (24) coincide com (8).

2.8 Resíduos

Após o ajustamento do modelo de Cox aos dados, surge a necessidade de analisar os resíduos do modelo. Esta é uma parte importante em qualquer análise regressão. A análise dos resíduos permite avaliar a validade dos pressupostos do modelo. Caso algum dos pressupostos não seja satisfeito, a interpretação dos resultados pode levar a conclusões erradas. No caso do modelo de Cox, os resíduos permitem avaliar:

I. Proporcionalidade das funções de risco;

II. Relação log-linear entre a variável resposta e uma covariável;

III. Existência de valores aberrantes (*outliers*) e de observações influentes.

Os resíduos, no contexto da análise de sobrevivência, não podem simplesmente ser calculados da forma habitual. Têm sido propostos vários tipos de resíduos para o modelo de Cox que permitem diagnosticar cada um dos aspetos mencionados acima:

Resíduos de Schoenfeld: permitem testar a proporcionalidade global das funções de risco e a proporcionalidade para cada covariável. Desta forma é possível verificar se o seu efeito é constante ao longo do tempo;

Resíduos martingala: tanto são usados para investigar a forma funcional de uma covariável como para identificar (*outliers*);

Resíduos *deviance*: são usados para identificar valores (*outliers*).

Resíduos *score*: são usados para identificar observações influentes.

Os resíduos martingala e *score* fazem parte de uma classe de resíduos obtidos por transformação de martingalas que foi proposta por Barlow e Prentice (1988). Por seu lado, os resíduos de Schoenfeld (1982) são um caso particular dos resíduos *score*.

2.8.1 Resíduos de Schoenfeld

Para a definição dos resíduos de Schoenfeld irá assumir-se que os tempos de vida são distintos. Sejam m tempos de vida distintos, $t_1 < t_2 < \cdots < t_m$, observados numa amostra de dimensão $n \geq m$. A cada instante t_j corresponde um conjunto R_j de indivíduos em risco. Seja $\mathbf{z}_j$ o vetor de p covariáveis associado ao indivíduo j, $j = 1, \cdots, n$. No modelo de Cox, para um indivíduo j com tempo de vida t_j, a variável aleatória z_{jk} apresenta o seguinte valor esperado condicional a R_j

$$\mathrm{E}(z_{jk}|R_j) = \frac{\sum\limits_{i \in R_j} z_{ik} \mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_i}}{\sum\limits_{i \in R_j} \mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_i}}.$$

O estimador de máxima verosimilhança parcial $\hat{\boldsymbol{\beta}}$ é solução de,

$$\mu_k(\boldsymbol{\beta}) = \sum_{j=1}^{m} \left[z_{jk} - \frac{\sum\limits_{i \in R_j} z_{ik} \mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_i}}{\sum\limits_{i \in R_j} \mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_i}} \right] = 0 \quad \Longleftrightarrow \quad \sum_{j=1}^{m} [z_{jk} - \mathrm{E}(z_{jk}|R_j)] = 0.$$

Substituindo $\boldsymbol{\beta}$ por $\hat{\boldsymbol{\beta}}$ em $\mathrm{E}(z_{jk}|R_j)$ obtém-se $\hat{\mathrm{E}}(z_{jk}|R_j)$. Para o indivíduo j, Schoenfeld (1982) definiu o vetor de resíduos como sendo um vetor $\hat{\mathbf{r}}_j = (\hat{r}_{j1}, \cdots, \hat{r}_{jp})'$ em que

$$\hat{r}_{jk} = z_{jk} - \hat{\mathrm{E}}(z_{jk}|R_j) \tag{26}$$

Assim, o resíduo de Schoenfeld associado à covariável z_k representa a diferença entre o valor observado da covariável z_k para o indivíduo j e a média ponderada dos valores de z_k observados para todos os indivíduos em risco no instante t_j, sendo os pesos dados por $\exp(\boldsymbol{\beta}'\mathbf{z}_i)$. Um valor absoluto de $\hat{r}_{jk}$ elevado

indica que, entre os indivíduos que estavam em risco no instante t_j, era pouco provável que a morte ocorresse ao indivíduo com valor da covariável z_k igual a z_{jk}. Estes resíduos apenas são calculados em instantes em que ocorrem acontecimentos.

Grambsch e Therneau (1994) propuseram uma transformação dos resíduos de Schoenfeld. Estes novos resíduos são designados por resíduos de Schoenfeld ponderados e têm uma maior capacidade de diagnóstico do que os resíduos em (26).

Seja $\hat{\mathbf{r}}_j = (\hat{r}_{j1}, \hat{r}_{j2}, \cdots, \hat{r}_{jp})$ o vetor de resíduos de Schoenfeld para o indivíduo j e $\widehat{\text{var}}(\hat{\mathbf{r}}_j)$ a estimativa da matriz de covariância de $\hat{\mathbf{r}}_j$. O vetor de resíduos de Schoenfeld ponderados é dado por

$$\hat{\mathbf{r}}_j^* = [\widehat{\text{var}}(\hat{\mathbf{r}}_j)]^{-1}\hat{\mathbf{r}}_j. \tag{27}$$

A covariância entre $\hat{\mathbf{r}}_k$ e $\hat{\mathbf{r}}_l$ encontra-se na linha k e na coluna l da matriz de covariância e é dada por,

$$\widehat{\text{var}}(\hat{\mathbf{r}}_j)_{kl} = \sum_{i \in R_j} \frac{e^{\hat{\boldsymbol{\beta}}'\mathbf{z}_i}}{\sum\limits_{h \in R_j} e^{\hat{\boldsymbol{\beta}}'\mathbf{z}_h}} \left(z_{jk} - \frac{\sum\limits_{h \in R_i} z_{hk} e^{\hat{\boldsymbol{\beta}}'\mathbf{z}_h}}{\sum\limits_{h \in R_i} e^{\hat{\boldsymbol{\beta}}'\mathbf{z}_h}} \right) \left(z_{jl} - \frac{\sum\limits_{h \in R_i} z_{hl} e^{\hat{\boldsymbol{\beta}}'\mathbf{z}_h}}{\sum\limits_{h \in R_i} e^{\hat{\boldsymbol{\beta}}'\mathbf{z}_h}} \right).$$

Dado o esforço computacional necessário para a realização destes cálculos, Grambsch e Therneau (1994) sugeriram uma aproximação para os resíduos (27). Esta sugestão foi baseada no facto da matriz $\widehat{\text{var}}(\hat{\mathbf{r}}_j)$ apresentar valores razoavelmente constantes ao longo do tempo de observação. Assim, o valor da sua inversa pode ser aproximado por

$$[\widehat{\text{var}}(\hat{\mathbf{r}}_j)]^{-1} = m\widehat{\text{var}}(\hat{\boldsymbol{\beta}}) \tag{28}$$

onde m representa o número de mortes observadas. Na prática, os resíduos de Schoenfeld ponderados são obtidos a partir de (27) e (28).

Considere-se então que o efeito da covariável z_k (fixa) pode variar ao longo do tempo e portanto poderá ser escrito como

$$\beta_k(t) = \beta_k + \gamma_k g_k(t).$$

Demonstra-se que o valor esperado no instante t do resíduo definido em (27) é aproximadamente igual à parte de $\beta_k(t)$ que varia com o tempo. De facto, Grambsch e Therneau (1994) provaram que $\gamma_k g_k(t) \approx \text{E}[\hat{r}_{jk}^*(t)]$, donde $\beta_k(t) \approx \hat{\beta}_k + \text{E}[\hat{r}_{jk}^*(t)]$, em que $\hat{\beta}_k$ é a estimativa de β_k no modelo de Cox ajustado aos dados.

Assim, a representação gráfica de $\hat{r}_{jk}^* + \hat{\beta}_k$ em função do tempo permite verificar se os resíduos apresentam uma forma sugestiva de não proporcionalidade, uma vez que se existir proporcionalidade dos riscos o gráfico não apresentará nenhum padrão definido. É comum considerar-se uma mudança da escala do tempo para que os resíduos fiquem espalhados pelo gráfico de forma mais homogénea, sendo assim mais fácil visualizar qualquer padrão. Por vezes, recorre-se à utilização do complementar da

estimativa de Kaplan-Meier $(1 - S_{KM})$, bem como do logaritmo do tempo. Os gráficos são geralmente complementados com a utilização de um suavizador como, por exemplo, o suavizador *lowess*, para facilitar a interpretação. Obtém-se assim informação sobre a forma de $\beta_k(t)$; por exemplo: uma linha horizontal sugere que o efeito de z_k é constante. Grambsch e Therneau (1994) aconselham que o gráfico dos resíduos seja complementado com a realização de um teste formal das hipóteses: $H_0 : \gamma_k = 0 \quad vs \quad H_a : \gamma_k \neq 0$, cuja estatística de teste, sob H_0, tem uma distribuição assintótica χ^2_1. Além deste teste para cada uma das covariáveis é também feito um teste global da hipótese de riscos proporcionais. Note-se que este é de facto um teste da correlação linear entre o tempo de vida e os resíduos.

2.8.2 Resíduos martingala

A teoria dos processos de contagem, como já foi referido anteriormente, é bastante útil na implementação do modelo de Cox. De seguida, serão apresentados alguns conceitos que envolvem processos de contagem, cuja definição se encontra no apêndice B.

Assuma-se que se está a seguir um indivíduo com vetor de covariáveis $\mathbf{z}$ e que $N(t)$ é uma função que toma o valor zero até imediatamente antes do instante em que o acontecimento ocorre e toma o valor um a partir daí. Esta função define um processo de contagem que indica a ocorrência ou não do acontecimento. $N(t)$ pode ser modelado em função de uma componente sistemática e de uma componente de erro, em que a primeira é a função de risco cumulativa associada ao modelo de Cox. Assim, tem-se $N(t) = \Lambda(t, \mathbf{z}, \boldsymbol{\beta}) + M(t)$ e, portanto, pode definir-se o resíduo martingala como

$$M(t) = N(t) - \Lambda(t, \mathbf{z}, \boldsymbol{\beta}). \tag{29}$$

Teoricamente (29) apresenta um valor para cada instante t, mas apenas se irá considerar o seu valor para cada indivíduo j no fim do período de *follow-up*.

Para cada indivíduo j com tempo de vida t_j define-se o resíduo martingala como sendo a diferença entre o número de acontecimentos ocorridos durante o tempo de observação e o número de acontecimentos esperados sob o modelo de Cox, que é dado por

$$M_j = N_j - \Lambda(t_j, \mathbf{z}_j, \boldsymbol{\beta}) = N_j - E_j. \tag{30}$$

Como não se conhece o verdadeiro valor de $\boldsymbol{\beta}$, então M_j é estimado por $\hat{M}_j$ em que se substituiu $\boldsymbol{\beta}$ por $\hat{\boldsymbol{\beta}}$. Quando $\hat{M}_j < 0$, o número de acontecimentos observados é menor do que o estimado pelo modelo e a sobrevivência está sobrestimada. As propriedades dos resíduos martingala são:

- ◇ O valor esperado para cada resíduo é zero quando se considera o verdadeiro valor do vetor de parâmetros desconhecidos $\boldsymbol{\beta}$: $\mathrm{E}(M_j) = 0$;

- ◇ O somatório dos resíduos estimados a partir de $\hat{\boldsymbol{\beta}}$ é igual a zero: $\sum \hat{M}_j = 0$;
- ◇ Os resíduos obtidos a partir do verdadeiro valor de $\boldsymbol{\beta}$ são não correlacionados: $\text{cov}(M_i, M_j) = 0$. No entanto, as estimativas destes resíduos, obtidas a partir de $\hat{\boldsymbol{\beta}}$, apresentam uma correlação negativa fraca: $\text{cov}(\hat{M}_i, \hat{M}_j) < 0$.

A partir de (30) pode dizer-se simplesmente que os resíduos martingala são estimados a partir da diferença $O-E$, entre o número observado de acontecimentos para um indivíduo e o número esperado. Este valor irá realçar os indivíduos mal ajustados pelo modelo de Cox. Estas situações podem ocorrer quando um indivíduo morre muito tarde, apesar de ter risco elevado de morte, ou quando morre muito cedo e as suas características observadas indicam um baixo risco de morte.

A análise dos resíduos martingala pode ser feita a partir de dois gráficos:

- ◦ Gráfico em que se representam os valores $\hat{M}_j$ no eixo das ordenadas e os valores estimados $\hat{\boldsymbol{\beta}}'\mathbf{z}_j$ no eixo das abcissas. Desta forma são realçados os indivíduos mal ajustados pelo modelo, que costumam ser encontrados entre os que têm tempos de vida muito longos ou muito curtos;
- ◦ Gráfico em que se representam, no eixo das ordenadas, os valores $\hat{M}_j$ estimados a partir de um modelo sem covariáveis ($\boldsymbol{\beta} = \mathbf{0}$) e no eixo das abcissas os valores de uma covariável contínua, juntamente com uma curva de suavização que sugere a forma funcional da covariável. Therneau *et al.* (1990) mostraram que se o modelo correto para uma covariável z_j é $\exp(\beta_j f(z_j))$ para uma dada função suave f, então a curva de suavização para z_j irá revelar, sob certas circunstâncias, a forma de f. Obviamente que, se a curva for linear, não é necessário fazer uma transformação dessa covariável.

2.8.3 Resíduos *deviance*

A distribuição dos resíduos martingala é bastante assimétrica, particularmente no caso em que apenas se observa a ocorrência de um acontecimento. Os resíduos *deviance*, propostos por Therneau *et al.* (1990), são obtidos pela normalização dos resíduos martingala. Estes resíduos são simetricamente distribuídos em torno de zero, sendo mais fácil a sua interpretação em relação à dos resíduos martingala. Para cada indivíduo i o resíduo *deviance* é dado por

$$D_i = \text{sinal}(\hat{M}_i)\sqrt{-2 \times \left(l_{i(modelo)} - l_{i(saturado)}\right)},$$

onde $\text{sinal}(\hat{M}_i)$ é o sinal do resíduo martingala; $l_{i(modelo)}$ e $l_{i(saturado)}$ são as log-verosimilhanças parciais do modelo considerado e o modelo saturado, respetivamente.

Na presença de uma baixa percentagem de censura, os resíduos D_i têm uma distribuição aproximadamente normal o que faz com que, nesta situação, sejam mais úteis na deteção de outliers do que os

resíduos martingala. Quando existe uma percentagem elevada de censura, observar-se no gráfico um grande número de pontos próximos de zero e, portanto, os resíduos já não têm a aparência de uma amostra aleatória normal. É frequente construírem-se três gráficos: um gráfico dos resíduos *versus* os valores preditos do modelo ou *versus* o índice da observação; e um gráfico quantil-quantil.

2.8.4 Resíduos *score*

Os resíduos *score* são também definidos usando a teoria das martingalas e foram propostos por Therneau *et al.* (1990). Estes resíduos quantificam o contributo de cada indivíduo para a estatística *score*, ou seja, a influência que cada indivíduo exerce na estimativa de $\boldsymbol{\beta}$. Permitem, portanto, avaliar a diferença ocorrida na estimativa de $\boldsymbol{\beta}$ quando um determinado indivíduo é eliminado da análise. Assim, para cada indivíduo j é calculada a diferença entre a estimativa de $\boldsymbol{\beta}$ obtida com e sem esse indivíduo: $\triangle\boldsymbol{\beta} = \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-j)}$.

Formalmente, tal como os resíduos de Schoenfeld, os resíduos *score* são obtidos a partir da derivada parcial da função log-verosimilhança em ordem a $\beta_k,\ k = 1, \cdots, p$. Para uma amostra de dimensão n com m tempos de vida tem-se

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_k} = \sum_{j=1}^{n} \left[\delta_j (z_{jk} - a_{jk}) + \mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_j} \sum_{t_r \leq t_j} \frac{(a_{rk} - z_{jk})\delta_r}{\sum\limits_{l \in R_r} \mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_l}} \right], \text{ com } a_{jk} = \frac{\sum_l z_{lk} \mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_l}}{\sum_l \mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_l}} \tag{31}$$

onde δ_j denota o estado do indivíduo j. Segundo esta formulação, o indivíduo j apenas contribui para a derivada (31) até ao instante t_j. Isto significa que, se a observação dos indivíduos terminasse neste instante, então a j-ésima componente da derivada não seria afetada. Assim, o resíduo *score* para o indivíduo $j\ , j = 1, \cdots, n$ e para a covariável $z_k\ , k = 1, \cdots, p$ é

$$\hat{r}_{S_{jk}} = \delta_j (z_{jk} - \hat{a}_{jk}) + \mathrm{e}^{\hat{\boldsymbol{\beta}}'\mathbf{z}_j} \sum_{t_r \leq t_j} \frac{(\hat{a}_{rk} - z_{jk})\delta_r}{\sum\limits_{l \in R_r} \mathrm{e}^{\hat{\boldsymbol{\beta}}'\mathbf{z}_l}}. \tag{32}$$

Os resíduos *score* são uma modificação dos resíduos $\hat{r}_{jk}$ de Schoenfeld (26), uma vez que

$$\hat{r}_{S_{jk}} = \hat{r}_{jk} + \mathrm{e}^{\hat{\boldsymbol{\beta}}'\mathbf{z}_j} \sum_{t_r \leq t_j} \frac{(\hat{a}_{rk} - z_{jk})\delta_r}{\sum\limits_{l \in R_r} \mathrm{e}^{\hat{\boldsymbol{\beta}}'\mathbf{z}_l}}.$$

Tal como acontece com os resíduos de Schoenfeld, também a soma dos resíduos *score* é igual a zero; no entanto, o resíduo *score* pode ser não nulo para indivíduos a que correspondem observações censuradas. Para cada covariável z_k são calculados os valores de $\hat{r}_{S_{jk}}$, sendo estes representados no eixo das ordenadas de um gráfico em que os valores de z_k são representados no eixo das abcissas. Os valores extremos são os que influenciam fortemente a estimativa de β_k. Para facilitar a visualização é habitual

representar no gráficos os resíduos $\hat{r}_{S_{jk}}$ ponderados pela estimativa do erro padrão de $\hat{\beta}_k$.
Alguns comentários sobre os resíduos *score*:

* São calculados, quer para tempos de vida, quer para tempos censurados, o que se torna vantajoso quando a percentagem de censura é grande;
* São úteis quando para cada indivíduo pode ocorrer mais do que um acontecimento, por permitirem uma estimação robusta da variância de $\hat{\boldsymbol{\beta}}$, como será referido no capítulo 3.

2.9 Extensões do modelo de Cox

A definição do modelo de Cox considerada até agora assumiu que as funções de riscos são proporcionais para todas as covariáveis nele incluídas, tendo estas valores fixos durante todo o período de seguimento. No entanto, por vezes é necessário modificar o modelo de Cox para que este se adapte a outras situações que surgem na prática.

2.9.1 Modelo de Cox estratificado

Por vezes existe a necessidade de estratificar o modelo de riscos proporcionais de Cox, seja porque o planeamento do estudo definiu uma estratificação dos indivíduos à priori; seja pela violação da proporcionalidade dos riscos. Numa estratificação procede-se à divisão da amostra de n indivíduos em S grupos disjuntos e define-se uma função de risco subjacente para cada grupo. A estimativa obtida para $\boldsymbol{\beta}$ é igual em todos os estratos, ou seja, considera-se que o efeito das covariáveis é igual em todos os estratos definidos.
Os grupos de indivíduos ou estratos são definidos a partir das classes de uma variável categórica designada por variável de estratificação. Também é possível obter a estratificação a partir de duas ou mais variáveis, em que cada estrato é definido por uma combinação única de valores das variáveis envolvidas. Por exemplo, se se considerarem duas covariáveis A e B com respetivos valores $\{a_1, a_2\}$ e $\{b_1, b_2, b_3\}$, os estratos incluídos na análise serão 6: a_1b_1, a_1b_2, a_1b_3, a_2b_1, a_2b_2 e a_2b_3. No entanto, é necessário ter em atenção que, na presença de um elevado número de estratos, as estimativas de $\boldsymbol{\beta}$ apresentam pouca precisão e os testes de hipóteses realizados sobre os parâmetros perdem potência.
A estratificação deve ser realizada usando uma variável para a qual a estimação do seu efeito seja secundária em relação às restantes variáveis. As variáveis de estratificação podem ser definidas pelo planeamento do estudo ou podem ser definidas a partir de análises anteriores. Um exemplo de uma variável de estratificação definida pelo planeamento de um estudo multicêntrico é o hospital onde o

doente foi internado. Isto deve-se ao facto de, por vezes, os hospitais terem diferentes populações de doentes ou diferentes padrões de referência ou *guidelines*. Nesta situação a estratificação desempenha um papel idêntico ao da análise de variância em ensaios clínicos planeados com blocos.
A estratificação pode ainda ser usada quando a condição de proporcionalidade dos riscos não é satisfeita para uma das covariáveis incluídas no modelo de Cox. Essa variável será usada na análise como variável de estratificação. Mas, ao contrário do que acontece no caso dos estudos multicêntricos em que no início do estudo se define a variável de estratificação, aqui a variável só assume esse papel quando se verifica a não proporcionalidade dos riscos. Desta forma deixa de ser possível estimar o efeito desta variável, apesar de se ter contornado o problema da não proporcionalidade dos riscos.

Formalmente, considere-se uma variável de estratificação com S categorias, $\mathbf{z}$ o vetor de covariáveis e $\boldsymbol{\beta}$ o vetor de parâmetros desconhecidos. A cada estrato s faz-se corresponder a função de risco,

$$\lambda_s(t, \mathbf{z}) = \lambda_{0s}(t)\mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}}, \quad s = 1, \cdots, S. \tag{33}$$

Assume-se no modelo anterior que as funções de risco são proporcionais entre indivíduos do mesmo estrato. Se se considerarem dois indivíduos de um mesmo estrato s, $1 \leq s \leq S$, com vetores de covariáveis $\mathbf{z}_1$ e $\mathbf{z}_2$, admite-se que a razão seguinte não depende de t,

$$\frac{\lambda_m(t, \mathbf{z}_1)}{\lambda_m(t, \mathbf{z}_2)} = \frac{\lambda_{0m}(t)\mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_1}}{\lambda_{0m}(t)\mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_2}} = \mathrm{e}^{[\boldsymbol{\beta}'(\mathbf{z}_1 - \mathbf{z}_2)]}$$

No caso em que os indivíduos não pertencem ao mesmo estrato pode ocorrer não proporcionalidade dos riscos, uma vez que as funções de risco subjacente $\lambda_{01}(t), \cdots, \lambda_{0S}(t)$ são funções arbitrárias que não estão relacionadas entre si e por isso podem ser diferentes. Os parâmetros desconhecidos β_j não dependem do estrato a que os indivíduos pertencem e por isso o efeito de cada uma das variáveis é igual para todos os estratos.
A estimação do vetor $\boldsymbol{\beta}$ é feita a partir da maximização da função de verosimilhança parcial que agora incorpora a informação relativa aos vários estratos, e é dada por

$$L_S(\boldsymbol{\beta}) = \prod_{s=1}^{S} L_s(\boldsymbol{\beta}) = \prod_{s=1}^{S} \prod_{j=1}^{J_s} \frac{\mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_{sj}}}{\sum\limits_{i \in R_{sj}} \mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_{si}}} \tag{34}$$

onde,

- ⋆ $L_s(\boldsymbol{\beta})$ representa a função de verosimilhança parcial correspondente ao estrato s;
- ⋆ $J_1, J_2, \cdots, J_S$ é o número de tempos de vida distintos em cada estrato;
- ⋆ R_{sj} indica o conjunto de indivíduos em risco no j-ésimo instante de morte do estrato s.

Neste contexto, a função de verosimilhança parcial para tempos de vida empatados é dada por

$$L_S(\boldsymbol{\beta}) = \prod_{s=1}^{S} \prod_{j=1}^{J_s} \frac{\mathrm{e}^{\boldsymbol{\beta}' \mathbf{a}_{sj}}}{\left[\sum_{i \in R_{sj}} \mathrm{e}^{\boldsymbol{\beta}' \mathbf{z}_{si}} \right]^{d_{sj}}} \tag{35}$$

com $\mathbf{a}_{sj}$ e d_{sj} a desempenharem papéis idênticos aos de $\mathbf{a}_j$ e d_j na função (24), mas agora em cada estrato s. A estimação de $\boldsymbol{\beta}$ e de $\lambda_0(t)$ é realizada como habitualmente. Os estimadores assim obtidos gozam das propriedades assintóticas referidas anteriormente, o que permite fazer inferência sobre os parâmetros. Para cada estrato s podem estimar-se $\Lambda_{0s}(t)$ e $S_{0s}(t)$. Note-se que a verificação da condição de proporcionalidade das funções de risco também pode ser feita recorrendo à construção de modelos de Cox estratificados.

2.9.2 Covariáveis dependentes do tempo

No modelo de Cox definido nas secções anteriores consideraram-se fixos os valores observados de cada covariável para cada indivíduo em estudo. Portanto, desde o instante inicial de observação até à ocorrência do acontecimento ou até à censura, o valor assumido por cada covariável é constante. De facto, existem variáveis cujo valor nunca se altera, como é o caso do género ou da cor dos olhos. No entanto, existem situações em que as variáveis assumem valores diferentes durante o tempo em observação como, por exemplo, a cessação tabágica, iniciação de uma atividade física regular ou o valor da tensão arterial. Apesar do valor destas variáveis poder sofrer alterações ao longo do período de observação, é possível e frequente que se considere apenas o valor observado no início do estudo, sendo uma opção do investigador não considerar as alterações que possam surgir. Um exemplo desta situação são os hábitos tabágicos; pode ser registado apenas se o indivíduo é fumador ou não no instante inicial, ignorando-se posteriormente se se dá uma cessação ou iniciação tabágica. Nesta secção, tem particular interesse considerar covariáveis cujo valor possa registar alterações no decurso do seguimento do indivíduo. Como é razoável admitir que o valor da função de risco estará mais dependente dos valores observados no decurso do estudo do que do valor observado no seu início, é importante considerar uma extensão do modelo de Cox que permita incluir este tipo de covariáveis. A opção de incluir estas covariáveis no modelo deve ser bem fundamentada, pois acarreta um aumento de complexidade na aplicação dos métodos de inferência.

Existem dois tipos de covariáveis dependentes do tempo: internas e externas. As primeiras referem-se a características específicas de cada indivíduo que requerem uma observação direta do indivíduo ao longo

do tempo, não sendo possível determinar os seus valores a partir de outros fatores. Exemplos destas características são a tensão arterial ou o nível de colesterol. As variáveis externas são aquelas que não obrigam a um contacto direto com o indivíduo para que se possa obter o seu valor num determinado instante. As variáveis ambientais, como a temperatura do ar ou a percentagem de humidade, são variáveis deste tipo. Os fatores controlados pelo investigador segundo um protocolo de estudo, como a dosagem de um fármaco a administrar, são também considerados variáveis externas. Existe ainda uma variável que alguns autores consideram uma variável externa: a idade. A idade é uma variável com trajetória determinística, o que significa que é possível calcular o seu valor em qualquer instante desde que se tenha registado a data de nascimento no início da observação, não sendo necessário observar diretamente o indivíduo para se determinar o seu valor.

Independentemente de se considerarem covariáveis dependentes do tempo internas ou externas, defina-se $\mathbf{z}_i'(t_j) = [z_{i1}(t_j), \cdots, z_{ip}(t_j)]$ como o vetor de covariáveis observadas para o indivíduo i no instante t_j. O valor t_j representa o tempo que decorreu desde o início da observação do indivíduo. O modelo de Cox em (6) passa a escrever-se na forma,

$$\lambda(t, \mathbf{z}_i(t)) = \lambda_0(t) \mathrm{e}^{\boldsymbol{\beta}' \mathbf{z}_i(t)} \qquad i = 1, \cdots, n. \tag{36}$$

A função de verosimilhança parcial (8) passa a ser, neste caso,

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{J} \frac{\mathrm{e}^{\boldsymbol{\beta}' \mathbf{z}_j(t_j)}}{\sum\limits_{i \in R_j} \mathrm{e}^{\boldsymbol{\beta}' \mathbf{z}_i(t_j)}} \tag{37}$$

É de notar que os modelos com covariáveis dependentes do tempo exigem uma construção cuidada da base de dados. Além disso, já não são considerados modelos de riscos proporcionais visto que

$$\frac{\lambda(t; \mathbf{z}_j(t))}{\lambda(t; \mathbf{z}_k(t))} = \frac{\lambda_0(t) \mathrm{e}^{\boldsymbol{\beta}' \mathbf{z}_j(t)}}{\lambda_0(t) \mathrm{e}^{\boldsymbol{\beta}' \mathbf{z}_k(t)}} = \mathrm{e}^{\boldsymbol{\beta}' [\mathbf{z}_j(t) - \mathbf{z}_k(t)]}, \text{ depende de } t.$$

Um dos propósitos da inclusão de variáveis dependentes do tempo num modelo de Cox, consiste na verificação da condição de proporcionalidade dos riscos. O teste desta hipótese para uma covariável z_1, cujos valores são fixos durante todo o período de *follow-up*, envolve a criação de uma covariável dependente do tempo, $z_2(t) = z_1 + g(t)$, e o ajustamento de um modelo de Cox de riscos proporcionais onde são incluídas z_1 e z_2, dado por

$$\lambda(t, z_1) = \lambda_0(t) \mathrm{e}^{[\beta_1 z_1 + \beta_2 (z_1 \times g(t))]},$$

em que $g(t)$ é uma função frequentemente definida como $g(t) = \ln t$. Se se compararem dois indivíduos

com valores distintos para z_1, $c_1 \neq c_2$, a razão das suas funções de risco

$$\frac{\lambda(t; z_1 = c_1)}{\lambda(t; z_1 = c_2)} = \mathrm{e}^{\beta_1[c_1 - c_2] + \beta_2 g(t)[c_1 - c_2]}$$

dependerá de t se e só se $\beta_2 \neq 0$. Assim, sob estas condições, o teste das hipóteses $H_0 : \beta_2 = 0 \quad vs \quad H_a : \beta_2 \neq 0$ é um teste da proporcionalidade de riscos para z_1.

3
Modelos para acontecimentos múltiplos

3.1 Introdução

O modelo de Cox é adequado para situações em que os tempos correspondentes aos indivíduos em estudo são independentes entre si e em que para cada indivíduo apenas se pode observar uma ocorrência do acontecimento. Nestes casos, uma vez ocorrido o acontecimento, o indivíduo deixa de estar em risco e sai do estudo. Exemplos deste tipo de acontecimentos são a morte, a menopausa ou o diagnóstico de uma doença crónica.

Existem, no entanto, situações em que os tempos observados estão correlacionados, e nesse caso a utilização direta do modelo de riscos proporcionais de Cox não é adequada, mesmo quando se usa a formulação dos processos de contagem. De facto, a inexistência de independência entre as observações e a manutenção dos indivíduos no grupo de risco depois do acontecimento ocorrer, levaria à obtenção de estimativas pouco fiáveis.

A correlação entre os tempos de vida pode ocorrer de duas formas:

1. Quando para cada indivíduo se observam vários tempos de vida, associados à ocorrência de múltiplos acontecimentos, o que é comum em contextos clínicos como a observação de diversos enfartes do miocárdio ou efeitos secundários de medicamentos;
2. Quando os indivíduos estão de alguma forma agrupados, quer natural, quer artificialmente. Desta forma, os tempos de vida estão correlacionados entre os indivíduos do mesmo grupo. São exemplos, os tempos até um determinado acontecimento, observados em indivíduos pertencentes à mesma família, ou em alunos de uma escola que pertencem à mesma turma. Não se leva em consideração a ordem pela qual ocorrem os acontecimentos para os indivíduos do mesmo grupo.

Quando se observam acontecimentos múltiplos, estes podem ser do mesmo tipo ou de tipos diferentes. Assim sendo, pode observar-se a recorrência ou repetição do mesmo acontecimento (enfarte), ou pode observar-se uma ocorrência de vários acontecimentos de naturezas diferentes (insuficiência renal, insuficiência respiratória ou morte). Além do tipo de acontecimentos, é também importante definir se a ordem pela qual os acontecimentos ocorrem deve ser levada em conta ou se deve ser ignorada. É frequente, por exemplo, nos casos em que se observa o tempo até à ocorrência de diversas complicações de uma doença, considerar-se a ordem pela qual estas ocorrem, uma vez que a condição do doente se pode ir agravando a cada complicação observada.

3.2 Modelos alternativos

Existem várias abordagens para a modelação de acontecimentos múltiplos. Uma das possibilidades é a consideração de um modelo de regressão cuja variável dependente segue uma distribuição de Poisson, sendo esta variável definida como o número de acontecimentos ocorridos durante o tempo de observação do indivíduo. No entanto, esta abordagem apresenta algumas limitações: não é possível distinguir um indivíduo com k acontecimentos ocorridos num período de 10 dias de outro em que o mesmo número de acontecimentos ocorreu em 150 dias, assim como não é tido em conta se os acontecimentos são diferentes ou não.

Outra abordagem consiste na utilização de modelos de efeitos aleatórios também conhecidos na área da análise de sobrevivência por modelos com fragilidade. Nestes, cada indivíduo está associado a um nível de agregação de vários acontecimentos, que condicionados ao efeito aleatório são considerados independentes. Embora seja uma metodologia cada vez mais usada, apenas se aplica a acontecimentos da mesma natureza.

Finalmente, pode ainda recorrer-se a uma metodologia diferente das já mencionadas, que consiste na modelação dos tempos de vida múltiplos por intermédio de modelos ditos marginais. Estes são extensões do modelo de Cox que permitem lidar com diversas situações em que se observam acontecimentos múltiplos. A partir dos modelos marginais é possível estimar os parâmetros sem que a estrutura de dependência dos tempos de vida seja considerada, utilizando o modelo de Cox usual. Posteriormente, é determinado um estimador robusto da matriz de covariância que permite introduzir a correção necessária, dada a presença de tempos de vida correlacionados.

Os modelos marginais podem dividir-se em dois grandes grupos, de acordo com a existência ou não de uma estrutura de ordenação dos acontecimentos. Os modelos de acontecimentos não ordenados mais usuais são o modelo de riscos paralelos e o modelo de Lee, Wei & Amato (LWA). Por seu turno, os modelos de acontecimentos ordenados mais comuns são: o modelo com risco condicional ou modelo de Prentice, Williams & Peterson (PWP), o modelo de incrementos independentes ou modelo de Andersen-Gill (AG) e o modelo com risco concomitante ou modelo de Wei, Lin & Wesseifeld (WLW). No caso em que se têm acontecimentos ordenados, os tempos de vida observados para cada indivíduo têm obrigatoriamente uma ordem, que é dada tanto pela definição das datas em que se inicia a observação para cada acontecimento, como pela ordem do estrato de risco correspondente a cada acontecimento. Este trabalho debruçar-se-á apenas sobre os modelos PWP, AG, WLW e LWA. Antes de se definirem formalmente estes quatro modelos, é importante conhecer alguns conceitos básicos que fazem com estes difiram entre si e também introduzir alguma notação.

3.3 Conceitos básicos

Os quatro modelos referidos na secção anterior diferem essencialmente em relação a quatro componentes: intervalo de risco, função de risco subjacente, grupo de indivíduos em risco e estrutura de dependência entre os acontecimentos.

No que respeita ao intervalo de risco, este define-se como o intervalo de tempo durante o qual se pode observar a ocorrência do acontecimento. A sua definição costuma ser feita durante a construção da base de dados. São possíveis três formulações diferentes: tempo total (*total time*), tempo por intervalos (*gap time*) e processo de contagem (*counting process*). No caso do tempo total, o instante inicial a partir do qual se determina o tempo até cada um dos acontecimentos é fixo e igual para todos os acontecimentos; habitualmente, é o instante em que o indivíduo entra no estudo. O tempo por intervalos, tal como o nome indica, divide o tempo de observação do indivíduo em intervalos, desde o instante de entrada até ao instante de saída do grupo de indivíduos em risco, sendo que, depois de cada acontecimento, o tempo recomeça a contar. A definição por processo de contagem apresenta a mesma escala de tempo definida pelo tempo total, no entanto, aqui existe a possibilidade do acontecimento ter uma certa duração, não sendo portanto instantâneo. Assim sendo, durante esse período de tempo o indivíduo não está em risco.

A função de risco subjacente é uma função não especificada que pode ser comum a todos os acontecimentos, independentemente da ordem pela qual ocorrem, ou pode variar segundo essa ordem; se um acontecimento ocorrer k vezes, então existirão k funções de risco subjacente.

Quanto à definição do conjunto de indivíduos em risco, são três as possibilidades para a sua definição: não restritivo, semi-restritivo e restritivo. Num conjunto não restritivo são considerados em risco todos os indivíduos, independentemente do número de acontecimentos já ocorridos, tendo-se uma função de risco subjacente comum a todos os acontecimentos. O grupo de indivíduos em risco para o acontecimento de ordem s inclui, no caso semi-restritivo, os indivíduos para os quais ocorreram no máximo $(s-1)$ acontecimentos e no caso restritivo, inclui apenas os indivíduos com exatamente $(s-1)$ acontecimentos. Nestes dois casos, consideram-se funções de risco específicas para cada acontecimento.

Finalmente, a estrutura de dependência entre os acontecimentos pode ser marginal, condicional ou definida segundo efeitos aleatórios. No caso marginal, os indivíduos estão simultaneamente em risco para todos os acontecimentos e a ocorrência de cada um deles não está dependente da ocorrência de algum dos outros. No caso condicional, a dependência já se verifica, na medida em que os indivíduos não podem estar em risco para a ocorrência de um acontecimento, sem que o anterior tenha ocorrido. Os modelos com efeitos aleatórios, também designados por modelos com fragilidade, incluem uma variável aleatória que representa a dependência existente entre os tempos de vida múltiplos, ou melhor, entre os indivíduos da mesma família ou grupo.

Do ponto de vista da estimação, os modelos AG, PWP, WLW e LWA são todos classificados como modelos marginais, embora no que diz respeito à estrutura de dependência, os modelos AG e PWP sejam condicionais e os modelos WLW e LWA sejam marginais.

3.4 Notação

Dada uma amostra de dimensão n, em que a cada indivíduo i $(1 = 1, \cdots, n)$ correspondem os tempos de vida T_{ki} e os tempos censurados C_{ki} para um acontecimento k $(k = 1, \cdots, K)$, o tempo de observação é dado por $X_{ki} = \min(T_{ki}, C_{ki})$. O vetor de p covariáveis para o indivíduo i e o acontecimento k é $\mathbf{z}_{ki}(t) = (z_{ki1}(t), \cdots, z_{kip}(t))'$ e $\mathbf{z}_i(t) = (\mathbf{z}'_{1i}(t), \cdots, \mathbf{z}'_{ki}(t))$ é o vetor de covariáveis observado para o i-ésimo indivíduo. Os vetores $\mathbf{X}_i = (X_{1i}, \cdots, X_{Ki})'$ e $\mathbf{C}_i = (C_{1i}, \cdots, C_{Ki})'$ são condicionalmente independentes dado $\mathbf{z}_{ki}(t)$. Seja $G_{ki} = X_{ki} - X_{k-1,i}$ o intervalo de tempo (*gap time*) com $X_{0i} = 0$.
Seja I(.) a variável indicatriz, tal que $\mathrm{I}(E) = 1$ quando E é verdadeira $\mathrm{I}(E) = 0$ quando é falsa. O estado do indivíduo é dado por $\delta_{ki} = \mathrm{I}(T_{ki} \leq C_{ki})$.
Denote-se $\lambda_{ki}(t)$ como a função de risco para o indivíduo i e o acontecimento k, $\lambda_0(t)$ como a função de risco subjacente não especificada comum a todos os acontecimentos e $\lambda_{k0}(t)$ a função de risco subjacente específica do acontecimento k. Finalmente, o vetor global de parâmetros de regressão desconhecidos é $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)'$. Já o vetor de parâmetros específico do acontecimento k é $\boldsymbol{\beta}_k = (\beta_{1k}, \cdots, \beta_{pk})'$.

3.5 Modelo PWP

3.5.1 Introdução

Uma das primeiras extensões do modelo de Cox (1972) para acontecimentos múltiplos foi desenvolvida por Prentice, Williams e Peterson (1981). Este modelo ficou conhecido por modelo PWP. Os autores consideraram a metodologia apresentada como uma generalização dos métodos de análise de sobrevivência já existentes, passando a ser possível modelar a função de risco para além do primeiro acontecimento.
Podem ser considerados dois modelos PWP que diferem apenas na definição da escala de tempo. O modelo PWP *counting process* (PWP-CP) define a escala de tempo fazendo uso de processos de contagem, enquanto o modelo PWP *gap time* (PWP-GT) considera o tempo definido por intervalos. O modelo PWP-CP foi inicialmente designado por modelo PWP com tempo total por Prentice *et al.* (1981), mas esta designação não é atualmente utilizada.

A definição do intervalo de risco destes dois modelos é mais fácil mediante a apresentação de um exemplo. Considere-se um indivíduo para o qual foram observados acontecimentos nos instantes 2, 5 e 13. Segundo a formulação dos processos de contagem, e considerando que se iniciou a observação no instante 0, o indivíduo está em risco para o primeiro acontecimento no intervalo $(0, 2]$; para o segundo acontecimento no intervalo $(2; 5]$; e para o terceiro acontecimento no intervalo $(5, 13]$. No caso do modelo PWP-GT os intervalos de risco são respetivamente, $(0, 2]$, $(0, 3]$ e $(0, 8]$. Nesta última formulação, o tempo define-se como o tempo desde o último acontecimento, ou seja, após cada acontecimento o relógio reinicia a contagem. É de notar que nas duas formulações os indivíduos estão em risco em intervalos de tempo com a mesma amplitude, mas considerando escalas de tempo diferentes. A utilização de processos de contagem permite que existam intervalos de tempo em que o indivíduo não está em risco, o que acontece quando o acontecimento tem uma duração mensurável. Apenas a título ilustrativo, se se considerar a definição da escala de tempo segundo o tempo total, os intervalos de risco para o exemplo anterior são: $(0, 2]$, $(0, 5]$ e $(0, 13]$. Apesar da escala de tempo aqui considerada ser a mesma que é usada no modelo PWP-GT, a amplitude destes intervalos de risco é diferente, o que pode influenciar significativamente as estimativas dos riscos relativos. Esta última formulação é considerada para modelos marginais, enquanto as outras duas se aplicam em modelos condicionais, daí o modelo PWP-CP ter deixado de ser designado por PWP com tempo total. As definições das escalas de tempo têm impacto na interpretação dos riscos relativos. No caso do modelo PWP-CP, estuda-se o efeito de uma covariável no tempo desde o início da observação, enquanto no caso do modelo PWP-GT, esse efeito refere-se ao tempo desde o último acontecimento.

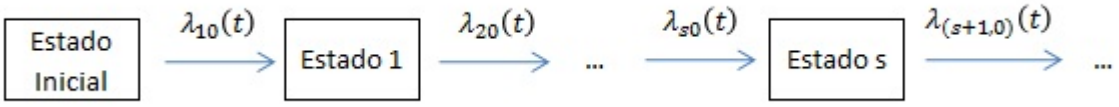
A função de risco subjacente é específica para cada acontecimento, porque se assume que após cada acontecimento o risco de sofrer o próximo se alterou em relação ao anterior, o que se traduz pela inclusão no modelo de estratos ordenados, um para cada acontecimento. Assim, para a k-ésima ocorrência do acontecimento, a função é representada por $\lambda_{k0}(t)$.

Finalmente, o grupo de indivíduos em risco para cada acontecimento é restritivo, por se considerar que os indivíduos apenas estão em risco para o acontecimento de ordem s se já tiverem sofrido o de ordem $(s-1)$. Por essa razão, o modelo PWP é também conhecido por modelo de acontecimentos ordenados com risco condicional, uma vez que o risco de ocorrência de cada acontecimento está condicionado pela ocorrência de acontecimentos anteriores.

É importante realçar que a dimensão do conjunto de indivíduos em risco tende a diminuir à medida que s aumenta, o que pode tornar as estimativas pouco precisas. Assim, ao analisar os dados, o número máximo de acontecimentos considerados deve ser escolhido tendo em conta esse facto.

3.5.2 Definição

Independentemente da formulação do intervalo de risco, o modelo PWP pode ser aplicado a situações como a representada a seguir:



Formalmente, considere-se uma amostra de n indivíduos e um vetor de p covariáveis possivelmente dependentes do tempo $\mathbf{z}(t) = (z_1(t), \cdots, z_p(t))'$. Usando a formulação de Cox (1972), a função de risco para o modelo PWP pode ser escrita como o produto de uma função arbitrária do tempo e uma função exponencial das covariáveis. Embora sejam possíveis outras escalas de tempo, apenas são consideradas as duas mais usuais na definição da função de risco subjacente para o acontecimento s. Assim, para um instante t, $\lambda_{s0}(t)$ é uma função do tempo decorrido desde o início do estudo até esse instante, ou é uma função do tempo desde o último acontecimento ocorrido, que é dado por $t - t_{i,s-1}$.
Para um indivíduo i, os dois modelos semi-paramétricos PWP têm funções de risco dadas por

$$\lambda_{si}(t) = \lambda_{s0}(t)\mathrm{e}^{\boldsymbol{\beta}_s'\mathbf{z}_{si}(t)} \quad \text{(modelo PWP-CP)}, \tag{38}$$

$$\lambda_{si}(t) = \lambda_{s0}(t - t_{i,s-1})\mathrm{e}^{\boldsymbol{\beta}_s'\mathbf{z}_{si}(t)} \quad \text{(modelo PWP-GT)}, \tag{39}$$

onde $\lambda_{s0}(.) \geq 0$ são funções de risco subjacentes completamente arbitrárias, $\boldsymbol{\beta}_s$ é o vetor de parâmetros de regressão específico de cada estrato e a variável de estratificação s pode variar com o tempo para cada indivíduo. De facto, o indivíduo passa para o estrato s imediatamente após a $(s-1)$-ésima ocorrência do acontecimento, e permanece em s até à s-ésima ocorrência do acontecimento ou até à censura.
É importante notar que, além de $\mathbf{z}(t)$ depender do tempo, pode também depender do número de acontecimentos já ocorridos. Para o indivíduo j em risco para o acontecimento de ordem s pode ter-se $\mathbf{z}_{sj}(t) = (z_{sj1}(t), \cdots, z_{sjp}(t))'$. À semelhança do modelo de Cox usual, o valor $\mathrm{e}^{\boldsymbol{\beta}_s'\mathbf{z}_s(t)}$ representa, nos dois modelos, o risco relativo associado a $\mathbf{z}_s(t)$.
Em relação à organização da base de dados, cada entrada refere-se a um tempo. É importante que a cada tempo se faça corresponder um indivíduo, o instante de entrada no conjunto de indivíduos em risco, o instante em que ocorreu o acontecimento ou a saída do estudo, a indicação se o tempo é um tempo de vida observado ou censurado e finalmente a ordem pela qual o tempo foi observado naquele indivíduo. Se se considerar novamente o exemplo dado anteriormente e se definir o instante 25 como o instante em que a observação do indivíduo termina, para os dois modelos serão registados quatro

instantes iniciais e quatro instantes finais. Para o modelo PWP-CP, os instantes iniciais são: 0, 2, 5 e 13; os instantes finais são 2, 5, 13 e 25. Para o modelo PWP-GT, os instantes iniciais são todos iguais a zero e os instantes finais são 2, 3, 8 e 12. Neste último caso, como os instantes iniciais são todos zero, alternativamente podem registar-se as amplitudes dos intervalos de tempo, ou seja, 2, 3, 8 e 12.

3.5.3 Função de verosimilhança parcial

Para fazer inferência em qualquer um dos modelos é necessário construir uma função de verosimilhança parcial. Podem obter-se estimadores dos vetores de parâmetros $\boldsymbol{\beta}_s$ específicos de cada estrato s ou obter um estimador global do vetor de parâmetros $\boldsymbol{\beta}$. O estimador global $\hat{\boldsymbol{\beta}}$ é obtido ajustando o modelo considerando um único vetor de covariáveis para o indivíduo j, $\mathbf{z}_j(t) = (z_{j1}(t), \cdots, z_{jp}(t))'$, ignorando os estratos. Genericamente, o estimador $\hat{\boldsymbol{\beta}}_s, s \geq 1$ é obtido ajustando um modelo com o vetor de covariáveis específico do estrato s, $\mathbf{z}_j(t) = (\mathbf{0}, \cdots, \mathbf{z}_{sj}(t), \cdots, \mathbf{0})'$, para o indivíduo j para o qual ocorreram os $s-1$ acontecimentos.
Considere-se em primeiro lugar o modelo (38). A função de verosimilhança parcial é

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \prod_{s \geq 1} \left(\frac{\mathrm{e}^{\boldsymbol{\beta}_s' \mathbf{z}_{si}(x_{si})}}{\sum_{j=1}^{n} \mathrm{Y}_{sj}(x_{si}) \mathrm{e}^{\boldsymbol{\beta}_s' \mathbf{z}_{si}(x_{si})}} \right)^{\delta_{si}}, \mathrm{Y}_{sj}(t) = \mathrm{I}(x_{s-1,j} < t \leq x_{sj}). \tag{40}$$

A função de verosimilhança parcial para o modelo (39) é

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \prod_{s \geq 1} \left(\frac{\mathrm{e}^{\boldsymbol{\beta}_s' \mathbf{z}_{si}(x_{(s-1,i)} + g_{si})}}{\sum_{j=1}^{n} \mathrm{Y}_{sj}(x_{si}) \mathrm{e}^{\boldsymbol{\beta}_s' \mathbf{z}_{si}(x_{(s-1,i)} + g_{si})}} \right)^{\delta_{si}}, \mathrm{Y}_{sj}(t) = \mathrm{I}(g_{sj} > t). \tag{41}$$

As propriedades assintóticas dos estimadores dos parâmetros do modelo referidas em Cox (1975) também são válidas para estes dois modelos. No entanto, é necessário ter em atenção o número de indivíduos e o número de acontecimentos ocorridos por estrato, especialmente quando se estima $\boldsymbol{\beta}_s, s \geq 1$.

3.6 Modelo AG

3.6.1 Introdução

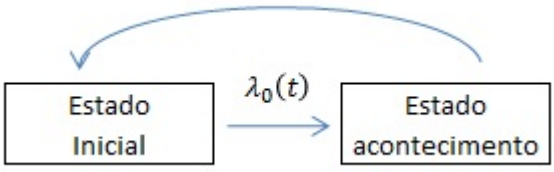
O modelo AG ou modelo de incrementos independentes foi proposto por Andersen e Gill (1982) e surge na sequência do modelo PWP. Dos quatros modelos referidos este é o mais simples, mas também o que apresenta pressupostos mais fortes, nomeadamente a independência dos tempos entre acontecimentos de um mesmo indivíduo. Tal como o modelo PWP, aplica-se a situações em que ocorrem acontecimentos

do mesmo tipo. No entanto, a ocorrência de um acontecimento não está dependente do número de acontecimentos ocorridos anteriormente, nem do tempo decorrido desde o último acontecimento, ou seja, as características do indivíduo não se alteram com a ocorrência de um acontecimento. Assume-se que, para qualquer indivíduo, o risco subjacente $\lambda_0(t)$ é comum a todos os acontecimentos. O modelo AG aplica-se, por exemplo, à ocorrência de infeções urinárias, desde que a recuperação do episódio não deixe sequelas, ou à gravidez, desde que o período do estudo seja suficientemente curto para que a idade não afete a fertilidade e o método de contraceção não se altere.

Tal como o modelo PWP, o intervalo de risco é definido por intermédio de processos de contagem, devendo definir-se como já foi referido anteriormente. Desta forma, cada indivíduo é contabilizado apenas uma vez no conjunto de risco para um dado acontecimento. É importante realçar que, apesar de se poder observar a ocorrência de acontecimentos múltiplos, estes nunca podem ocorrer em simultâneo para o mesmo indivíduo. O conjunto de indivíduos em risco para cada acontecimento é não restritivo, tendo em conta a definição do intervalo de risco e da função de risco subjacente. Lin (1994) recomenda que o modelo AG seja usado quando apenas se pretende estimar a taxa global de ocorrências da mesma natureza e se tem uma proporção baixa de indivíduos com dois ou mais acontecimentos.

3.6.2 Definição

Esquematicamente, o modelo AG aplica-se a situações como a seguinte:



Formalmente, considere-se uma amostra de n indivíduos e um vetor de p covariáveis possivelmente dependentes do tempo $\mathbf{z}(t) = (z_1(t), \cdots, z_p(t))'$, para um indivíduo i, o modelo semi-paramétrico AG tem função de risco dada por

$$\lambda_{ki}(t) = \lambda_0(t) \mathrm{e}^{\boldsymbol{\beta}' \mathbf{z}_{ki}(t)}, \qquad t \geq 0, \tag{42}$$

onde $\lambda_0(.) \geq 0$ é uma função de risco subjacente completamente arbitrária e comum a todos os acontecimentos, assim como o vetor de parâmetros $\boldsymbol{\beta}$.

Como a função de risco subjacente é comum a todos os acontecimentos este modelo não é estratificado, ao contrário do que acontecia com o modelo PWP, uma vez que o indivíduo regressa sempre ao seu estado inicial. O modelo AG pode ser considerado um modelo PWP-CP não estratificado, sendo a construção da base de dados feita de igual maneira nos dois modelos.

A função de verosimilhança parcial é

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \prod_{k \geq 1} \left(\frac{\mathrm{e}^{\boldsymbol{\beta}' \mathbf{z}_{ki}(x_{ki})}}{\sum_{j=1}^{n} \sum_{l \geq 1} \mathrm{Y}_{lj}(x_{ki}) \mathrm{e}^{\boldsymbol{\beta}' \mathbf{z}_{lj}(x_{ki})}} \right)^{\delta_{ki}}, \mathrm{Y}_{lj}(t) = \mathrm{I}(x_{l-1,j} < t \leq x_{lj}) \tag{43}$$

As propriedades assintóticas dos estimadores dos parâmetros são as mesmas enunciadas em Cox (1975), desde que o pressuposto de tempos independentes seja válido.

A independência entre os tempos significa que, para o mesmo indivíduo, a ocorrência de um determinado acontecimento não está dependente dos acontecimentos que já ocorreram no passado, desde que as covariáveis incluídas no modelo expliquem as diferenças existentes entre os indivíduos. No entanto, tal é pouco verosímil na prática. Quando não existe independência, a matriz de covariância dos estimadores dos parâmetros é sobrestimada quando se recorre ao estimador usual (12), devendo neste caso ser usada a matriz robusta que será definida na secção 3.9. Este pressuposto pode ser testado comparando as duas estimativas da matriz de covariância, a usual (isto é, admitindo independência) e a robusta. Se a variância robusta for apenas um pouco maior do que a usual, então pode aplicar-se o modelo AG; caso contrário, a condição de tempos independentes não é satisfeita, e portanto o modelo deixa de ser um modelo de riscos proporcionais, devendo usar-se como alternativa o modelo PWP ou o modelo com efeitos aleatórios.

3.7 Modelo WLW

3.7.1 Introdução

Wei, Lin e Wesseifeld (1989) propuseram um modelo semi-paramétrico, conhecido como modelo WLW. Este modelo surge com o objetivo de analisar, de uma forma geral, os tempos até à ocorrência de diversos acontecimentos, podendo estes ser de naturezas diferentes ou não.

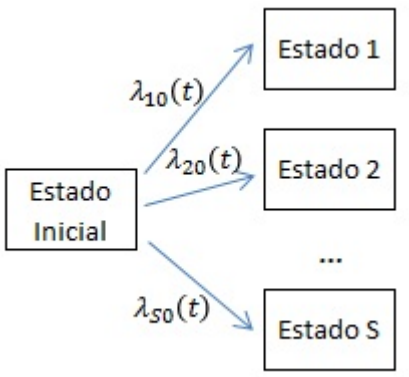
Até então, os modelos propostos impunham estruturas de dependência para acontecimentos observados no mesmo indivíduo e eram considerados como generalizações de metodologias já existentes, de modo a permitir que a função de risco continue a ser modelada após a ocorrência do primeiro acontecimento. Ao contrário dos modelos AG e PWP, o modelo WLW não impõe uma estrutura de dependência entre os tempos observados para o mesmo indivíduo, ou seja, a ocorrência de um acontecimento não está condicionada pela ocorrência de outro. Por esta razão, recorre-se ao modelo de Cox (1972) para modelar em separado o tempo até cada um dos acontecimentos e por isso o modelo WLW é também conhecido por modelo marginal.

Se se está interessado em observar S acontecimentos, o modelo WLW assume que cada indivíduo está

simultaneamente em risco de sofrer cada um deles desde o início da observação, podendo os acontecimentos ocorrer em simultâneo, daí que o conjunto de indivíduos em risco seja semi-restritivo. Por outro lado, cada indivíduo pode estar incluído simultaneamente nos S conjuntos de indivíduos em risco, o que apenas é permitido pelo modelo WLW. Sendo S o número máximo de acontecimentos observados para algum indivíduo, são definidos S estratos correspondentes aos S tipos de acontecimentos.
Neste caso, o intervalo de risco é definido desde o início da observação até à ocorrência de cada um dos acontecimentos. Para cada indivíduo observam-se S tempos, um para cada acontecimento. Caso não ocorram acontecimentos, registam-se S tempos de censura. Os indivíduos só deixam de estar em risco quando ocorreram todos os acontecimentos ou quando deixam de estar sob observação.
O modelo WLW considera um risco subjacente diferente consoante a ordem pela qual um acontecimento ocorre, daí serem consideradas S funções de risco subjacente, tantas quantos os acontecimentos a observar. Nesta análise, independentemente dos acontecimentos serem da mesma natureza ou não, o que interessa é a ordem pela qual ocorrem. Este modelo pode, por exemplo, ser usado para comparar dois esquemas terapêuticos no aparecimento de três efeitos secundários diferentes: A, B ou C. Para cada indivíduo são observados três tempos medidos desde o início da terapêutica. Considera-se que o risco de ocorrência do terceiro efeito secundário é diferente do risco de ocorrência do segundo, independentemente do seu tipo.

3.7.2 Definição

O esquema associado a este modelo é



Formalmente, considere-se uma amostra de n indivíduos e um vetor de p covariáveis possivelmente dependentes do tempo $\mathbf{z}(t) = (z_1(t), \cdots, z_p(t))'$, para um indivíduo i, a função de risco para o modelo WLW é definida por

$$\lambda_{si}(t) = \lambda_{s0}(t)\mathrm{e}^{\boldsymbol{\beta}_s'\mathbf{z}_{si}(t)}, \qquad t \geq 0, \tag{44}$$

onde $\lambda_{s0}(t)$ é a função de risco subjacente não especificada e $\boldsymbol{\beta}_s = (\beta_{1s}, \cdots, \beta_{ps})'$ é o vetor de parâmetros de regressão para o modelo marginal referente ao acontecimento $s = 1, \cdots, S$. Este é também um modelo estratificado pela ordem em que cada acontecimento ocorre.

Em relação à organização da base de dados, para cada indivíduo são registados S tempos correspondentes aos S acontecimentos cuja ocorrência se pretende avaliar. Neste caso, ao contrário do que acontece nos modelos PWP e AG, não se registam os instantes de entrada no conjunto de indivíduos em risco nem os instantes em que ocorreram os acontecimentos ou a saída do estudo. Aqui apenas se regista, para cada acontecimento de cada indivíduo, o tempo decorrido entre o instante inicial de observação e o instante final (no exemplo considerado, admitindo que existe um indivíduo para o qual se observam 5 acontecimentos e que o estudo durou 25 dias, registam-se os 5 tempos: 2, 5, 13, 25 e 25). A indicação se cada tempo se trata de um tempo de vida observado ou censurado e a ordem pela qual cada acontecimento ocorreu para aquele indivíduo é registado de forma análoga à dos dois modelos anteriores.

A função de verosimilhança parcial correspondente é

$$L_s(\boldsymbol{\beta}) = \prod_{i=1}^{n} \prod_{s=1}^{S} \left(\frac{\mathrm{e}^{\boldsymbol{\beta}' \mathbf{z}_{si}(x_{si})}}{\sum_{j=1}^{n} \mathrm{Y}_{sj}(x_{si}) \mathrm{e}^{\boldsymbol{\beta}' \mathbf{z}_{sj}(x_{si})}} \right)^{\delta_{si}}, \mathrm{Y}_{sj}(t) = \mathrm{I}(x_{sj} \geq t). \tag{45}$$

Tal como para o modelo PWP, o modelo WLW permite estimar $\boldsymbol{\beta}_s, (s = 1, \cdots, S)$ ou o vetor global de parâmetros $\boldsymbol{\beta}$. No entanto, para este último a forma de cálculo é distinta. No caso do modelo WLW o estimador de $\boldsymbol{\beta}$ é determinando pela média ponderada dos estimadores $\hat{\boldsymbol{\beta}}_s, (s = 1, \cdots, S)$, de modo que a correspondente média ponderada das variâncias robustas seja a menor possível.

3.8 Modelo LWA

3.8.1 Introdução

O modelo proposto por Lee, Wei e Amato (1992), conhecido por modelo LWA, surge numa perspetiva um pouco diferente dos anteriores. Enquanto nos outros três modelos se observavam vários acontecimentos para o mesmo indivíduo, o modelo LWA aplica-se quando os tempos estão agrupados segundo um elevado número de grupos independentes de pequena dimensão, não necessariamente a mesma.

Este modelo pode ser aplicado, por exemplo, quando se pretende estudar o tempo até à ocorrência de um determinado tipo de cancro numa amostra constituída por grupos de irmãos. Neste caso, seria observado um tempo por indivíduo, mas os tempos correspondentes aos irmãos estariam agrupados por estes partilharem uma parte do código genético. Outro exemplo de aplicação, que se pode encontrar

em Lee *et al.* (1992), refere-se ao tratamento de retinopatia diabética com sorbinil, em que se pretende modelar o tempo até à perda severa de visão. Como o tratamento afeta todo o organismo e como a perda de visão pode ocorrer apenas num dos olhos, observou-se o tempo até à ocorrência da perda de visão em separado para cada olho. Assim, os tempos foram agrupados por indivíduo, por se observarem os dois olhos.

O modelo LWA considera que o intervalo de risco é definido desde o início da observação até à ocorrência do acontecimento. Assume ainda um risco subjacente comum a todos os acontecimentos, com um conjunto de indivíduos em risco não restritivo, permitindo que um indivíduo esteja simultaneamente em risco para vários acontecimentos. Um indivíduo com k intervalos de risco pode ser incluído em k conjuntos de indivíduos em risco em qualquer instante em que está sob observação. Nenhum dos restantes modelos permite esta definição.

O modelo WLW considera um modelo LWA para cada acontecimento de natureza diferente, utilizando um conjunto de indivíduos em risco semi-restritivo. No entanto, estes dois modelos diferem ainda quanto à definição da função de risco subjacente; no caso do modelo WLW não é plausível considerar-se uma função comum a todos os acontecimentos, tanto no caso de acontecimentos de naturezas diferentes, como no caso de acontecimentos recorrentes do mesmo tipo.

3.8.2 Definição

O esquema associado a este modelo é idêntico ao do modelo WLW, mas neste caso a função subjacente é comum a todos os acontecimentos.

Tal como foi ilustrado pelos dois exemplos apresentados, podemos ter uma amostra de dimensão n em que os indivíduos estão agrupados em m grupos, de acordo com uma característica, sendo observado um acontecimento por indivíduo; ou podemos ter uma amostra de dimensão n em que para cada indivíduo se observam vários tempos de vida, e neste caso cada indivíduo representa um grupo. Sem perda de generalidade, na formalização do modelo é considerado o segundo caso.

Dada uma amostra de dimensão n, o k-ésimo tempo observado para o indivíduo i é t_{ki}. A dimensão dos n grupos é dada por $K_1, \cdots, K_n$, sendo as suas dimensões relativamente pequenas quando comparadas com n. Para um vetor de p covariáveis $\mathbf{z}(t) = (z_1(t), \cdots, z_p(t))'$, a função de risco do modelo LWA é

$$\lambda_{ki}(t) = \lambda_0(t) \mathrm{e}^{\boldsymbol{\beta}' \mathbf{z}_{ki}(t)}, \qquad t \geq 0. \tag{46}$$

Como $\lambda_0(t)$ é comum a todos os acontecimentos, o modelo LWA é um modelo não estratificado.

Em relação à organização da base de dados, para cada tempo é registado se se trata de um tempo de vida observado ou censurado e identifica-se o indivíduo correspondente. A definição do tempo é

idêntica à do modelo WLW e a função de verosimilhança parcial é

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n}\prod_{k=1}^{K_i}\left(\frac{\mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_{ki}(x_{ki})}}{\sum_{j=1}^{n}\sum_{l=1}^{K_i}\mathrm{Y}_{lj}(x_{ki})\mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_{lj}(x_{ki})}}\right)^{\delta_{ki}}, \mathrm{Y}_{lj}(t) = \mathrm{I}(x_{lj} \geq t) \tag{47}$$

Neste caso apenas tem interesse obter uma estimativa do vetor global de parâmetros.

3.9 Inferência

O aspeto mais importante a ter em conta na implementação dos modelos descritos anteriormente consiste na construção da base de dados a usar na análise. É necessário definir cuidadosamente os intervalos de risco (ou seja, os instantes de entrada e saída dos indivíduos no conjunto de indivíduos em risco), as variáveis de estratificação, bem como se o tempo observado é um tempo de vida ou de censura. Após a construção da base de dados deve fazer-se uma verificação exaustiva, principalmente quando as estruturas de sobrevivência são complexas. Embora este seja um passo bastante moroso, é indispensável para a obtenção de resultados fidedignos.

Os quatro modelos considerados utilizam o modelo de Cox para determinar o estimador de máxima verosimilhança parcial dos parâmetros, ignorando a estrutura de dependência dos tempos. Do ponto de vista computacional, podemos aplicar a metodologia dos modelos WLW e LWA nos modelos PWP e AG. Ao definir os conjuntos de risco para o tempo total usando $\mathrm{I}(x_{k-1,i} < t \leq x_{ki})$ em vez de $\mathrm{I}(x_{ki} \geq t)$, as funções de verosimilhança dos modelos LWA e WLW vão coincidir com as funções de verosimilhança dos modelos AG e PWP-CP, respetivamente. No caso do modelo PWP-GT, é necessário substituir $\mathrm{I}(x_{ki} \geq t)$ por $\mathrm{I}(x_{ki} - x_{k-1,i} \geq t)$ e $\mathbf{z}_{ki}(x_{ki})$ por $\mathbf{z}_{ki}(x_{k-1,i} + g_{ki})$, na função de verosimilhança de WLW. Por este motivo, do ponto de vista da estimação dos parâmetros, os modelos são todos modelos "marginais", diferindo apenas na definição dos conjuntos de indivíduos em risco.

Apesar de se estar perante tempos correlacionados, o estimador de máxima verosimilhança parcial obtido a partir do modelo de Cox é consistente e tem distribuição assintótica normal multivariada de dimensão p. No entanto, a matriz $\mathbf{I}^{-1}(\hat{\boldsymbol{\beta}})$ não é uma aproximação válida para a matriz de covariância por existir correlação entre os tempos e portanto, é necessário obter um estimador robusto para esta matriz. Também neste contexto se tira partido do facto dos modelos serem "marginais", já que os estimadores da matriz de covariância robusta dos modelos AG e PWP são os mesmos considerados para os modelos LWA e WLW, respetivamente.

Para uma amostra de dimensão n, $k = 1, \cdots, K$ e $r = 0, 1$ define-se,

$$\mathbf{S}_k^{(0)}(\boldsymbol{\beta}, t) = \sum_{j=1}^{n}\mathrm{Y}_{kj}(t)\mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_{kj}(t)}; \quad \mathbf{S}_k^{(1)}(\boldsymbol{\beta}, t) = \sum_{j=1}^{n}\mathrm{Y}_{kj}(t)\mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_{kj}(t)}\mathbf{z}_{kj}(t) \quad \text{e} \quad \overline{\mathbf{S}}^{(r)}(\boldsymbol{\beta}, t) = \sum_{k=1}^{K}\mathbf{S}_k^{(r)}(\boldsymbol{\beta}, t).$$

Considerando os modelos WLW e LWA com funções de risco definidas em (44) e (46), e funções de verosimilhança parcial (45) e (47), o estimador "*sandwich*" (estimador robusto para a matriz de covariância) é dado por

$$\mathbf{D}(\hat{\boldsymbol{\beta}}) = \mathbf{I}^{-1}(\hat{\boldsymbol{\beta}})\mathbf{V}(\hat{\boldsymbol{\beta}})\mathbf{I}^{-1}(\hat{\boldsymbol{\beta}}), \quad \text{em que} \quad \mathbf{V}(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{l=1}^{K}\mathbf{W}_{ki}(\hat{\boldsymbol{\beta}})\mathbf{W}_{li}(\hat{\boldsymbol{\beta}})' \tag{48}$$

O valor de $\mathbf{W}_{ki}$ é diferente conforme se considere um modelo estratificado ou não. Assim, para o modelo WLW, tem-se

$$\mathbf{W}_{ki}(\hat{\boldsymbol{\beta}}) = \delta_{ki}\left[\mathbf{z}_{ki}(x_{ki}) - \frac{\mathbf{S}_k^{(1)}(\hat{\boldsymbol{\beta}}, x_{ki})}{\mathbf{S}_k^{(0)}(\hat{\boldsymbol{\beta}}, x_{ki})}\right] - \sum_{j=1}^{n}\frac{\delta_{kj}\mathrm{Y}_{ki}(x_{kj})\mathrm{e}^{\hat{\boldsymbol{\beta}}'\mathbf{z}_{ki}(x_{kj})}}{\mathbf{S}_k^{(0)}(\hat{\boldsymbol{\beta}}, x_{kj})}\left[\mathbf{z}_{ki}(x_{kj}) - \frac{\mathbf{S}_k^{(1)}(\hat{\boldsymbol{\beta}}, x_{kj})}{\mathbf{S}_k^{(0)}(\hat{\boldsymbol{\beta}}, x_{kj})}\right].$$

enquanto para o modelo LWA,

$$\mathbf{W}_{ki}(\hat{\boldsymbol{\beta}}) = \delta_{ki}\left[\mathbf{z}_{ki}(x_{ki}) - \frac{\overline{\mathbf{S}}^{(1)}(\hat{\boldsymbol{\beta}}, x_{ki})}{\overline{\mathbf{S}}^{(0)}(\hat{\boldsymbol{\beta}}, x_{ki})}\right] - \sum_{j=1}^{n}\sum_{l=1}^{K}\frac{\delta_{lj}\mathrm{Y}_{ki}(x_{lj})\mathrm{e}^{\hat{\boldsymbol{\beta}}'\mathbf{z}_{ki}(x_{lj})}}{\overline{\mathbf{S}}^{(0)}(\hat{\boldsymbol{\beta}}, x_{kj})}\left[\mathbf{z}_{ki}(x_{lj}) - \frac{\overline{\mathbf{S}}^{(1)}(\hat{\boldsymbol{\beta}}, x_{lj})}{\overline{\mathbf{S}}^{(0)}(\hat{\boldsymbol{\beta}}, x_{lj})}\right].$$

Quando os tempos são correlacionados, é frequente que a estimativa da variância obtida a partir do estimador "*sandwich*" seja substancialmente superior à estimativa obtida admitindo que os tempos são independentes. No entanto, existem situações em que a estimativa robusta é inferior à estimativa usual, o que se deve ao facto de existir maior variabilidade entre os tempos observados para o mesmo indivíduo, do que entre os tempos observados para indivíduos diferentes, conforme referido por Kelly e Lim (2000).

Os três testes de hipóteses considerados em 2.5.1, utilizados para testar hipóteses em modelos aninhados ou testar hipóteses sobre o efeito individual de uma determinada covariável, podem ser usados, mas são mais conservadores do que os que utilizam a estimativa usual da variância. Por exemplo, no caso do teste de Wald, na estatística de teste apenas se substitui a variância usual pela variância robusta. No caso da determinação de intervalos de confiança, o procedimento é idêntico.

3.10 Algumas considerações

O modelo marginal WLW foi inicialmente proposto como uma alternativa ao modelo condicional PWP. Uma das vantagens consiste no facto da dimensão do conjunto de indivíduos em risco no caso do modelo WLW, ser maior que a dimensão para o modelo PWP, em todos os estratos. Como se sabe, no modelo PWP, a dimensão diminui com o aumento do número de acontecimentos ocorridos, o que introduz

pouca precisão nas estimativas associadas aos estratos correspondentes a um número de ocorrências mais elevado. Outra das vantagens do modelo WLW consiste na introdução de uma maior dependência entre as estimativas dos parâmetros de regressão específicos de cada estrato. De mais a mais, como a inferência é baseada nos modelos marginais, não é imposta nenhuma estrutura específica de dependência entre os tempos de vida.

Apesar destas vantagens, a utilização do modelo WLW para situações em que se observam acontecimentos recorrentes tem sido criticada. Uma das críticas prende-se com o facto de não ser definida uma ordem natural entre os acontecimentos recorrentes, podendo um indivíduo ser incluído no conjunto de indivíduos em risco para o k-ésimo acontecimento sem ainda ter sofrido o de ordem k-1. Esta formulação do conjunto de indivíduos em risco leva frequentemente a um efeito de arrastamento (*carry over*). Por exemplo, quando o objetivo da análise é a determinação do impacto de um tratamento no tempo de vida e este é significativo, espera-se que os indivíduos tratados tenham menos acontecimentos no mesmo período de tempo, comparativamente aos indivíduos não tratados. Um conjunto semi-restritivo inclui todos os indivíduos em cada estrato, e portanto, à medida que os acontecimentos vão ocorrendo o número de indivíduos tratados a que correspondem observações censuradas vai aumentando. Estas observações censuradas são comparadas com as dos indivíduos não tratados a quem está a ocorrer um acontecimento, o que exagera o efeito do tratamento nos últimos estratos.

Relativamente a conjuntos de indivíduos em risco não restritivos, estes apenas são adequados quando a função de risco subjacente não se altera com a ocorrência de cada acontecimento, o que é apropriado no caso do modelo AG, mas já não é no caso do modelo LWA. Este modelo tende a subestimar o efeito do tratamento por permitir que um indivíduo esteja em risco em vários instantes para o mesmo acontecimento.

É, portanto, essencial que se considere um conjunto de indivíduos em risco restritivo quando se lida com acontecimentos recorrentes para os quais se espera que o risco se altere após a ocorrência de cada acontecimento. Caso contrário, o efeito médio do tratamento não é estimado corretamente. Ao estimar os parâmetros do modelo específicos de cada acontecimento tem-se a vantagem de conseguir observar a forma como o efeito do tratamento se altera com cada acontecimento ocorrido. Embora o modelo PWP não produza um efeito *carry over* e seja possível escolher o número máximo de acontecimentos que podem ocorrer (de forma a que exista informação suficiente para estimar os parâmetros nos estratos a que corresponde um maior número de acontecimentos), a sua utilização também apresenta limitações. A interpretação dos parâmetros deste modelo é dificultada pela sua natureza condicional e, além disso, o facto da análise ser baseada num conjunto de indivíduos em risco restritivo viola a condição MCAR (*missing completely at random*), devido ao facto dos indivíduos que não sofreram o k-ésimo acontecimento serem excluídos da análise relativa ao $(k+1)$-ésimo acontecimento, conforme

referido por Cai e Schaubel (2004).

Portanto, para decidir qual a abordagem a usar na análise de acontecimentos recorrentes deve ter-se em conta o objetivo do estudo. Se apenas se estiver interessado na taxa global de recorrências do mesmo tipo, o modelo mais fácil de implementar é o AG, desde que se incluam covariáveis dependentes do tempo para definir a dependência. Se o objetivo recai nos intervalos de tempo e o risco se altera depois de cada ocorrência, então o modelo mais apropriado é o PWP. Se se pretender analisar o tempo desde o início da observação e se os acontecimentos puderem ocorrer simultaneamente, então deve recorrer-se a um modelo marginal, como o WLW.

4
Caso prático

4.1 Introdução

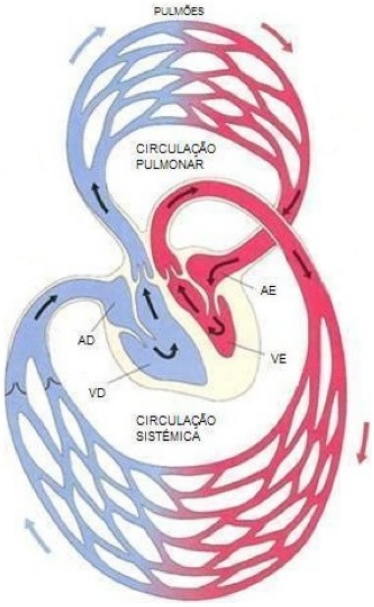
O enfarte agudo do miocárdio (EAM), constitui atualmente uma das principais causas de morte em todo o mundo. Apesar da mortalidade por EAM ter sofrido uma redução significativa desde meados dos anos sessenta do século XX, fruto da melhoria dos cuidados de saúde, da implementação das medidas terapêuticas e da mudança do estilo de vida, continua a ser um motivo de preocupação. O EAM resulta geralmente da morte do músculo cardíaco por obstrução de uma artéria coronária e consequente privação de oxigénio e nutrientes. O prognóstico é favorecido por uma menor área de enfarte e uma maior rapidez na obtenção de tratamento adequado. O EAM e a angina instável (AI) fazem parte das doenças coronárias, e em conjunto constituem as síndromes coronárias agudas (SCA). Durante uma hospitalização por SCA, além dos tratamentos realizados e da medicação administrada, é feito um acompanhamento do doente com vista a ajudá-lo a realizar uma mudança do estilo de vida, tentando-se prevenir que uma AI evolua para um EAM ou que se repita o EAM.
A partir de um registo contínuo, que é feito desde janeiro de 2004, de doentes hospitalizados por SCA no serviço de cardiologia dos Hospitais da Universidade de Coimbra, foram selecionados os que não tinham antecedentes de doença coronária, por forma a identificar os fatores que pudessem influenciar a ocorrência de novos enfartes.
A análise dos dados foi realizada recorrendo ao software estatístico R, versão 2.15.1.
Serão em seguida referidos alguns conceitos relacionados com o EAM, importantes para uma melhor compreensão do estudo (Gavina e Pinho, 2010).

4.2 O enfarte do miocárdio

4.2.1 Definição

Para definir o que é o EAM é essencial que se conheça a composição do coração e a sua função no organismo. De uma forma resumida, podemos descrever o coração como um órgão composto na sua maioria por tecido muscular, onde se encontram quatro cavidades: duas aurículas (AE e AD) e dois

ventrículos (VE e VD).

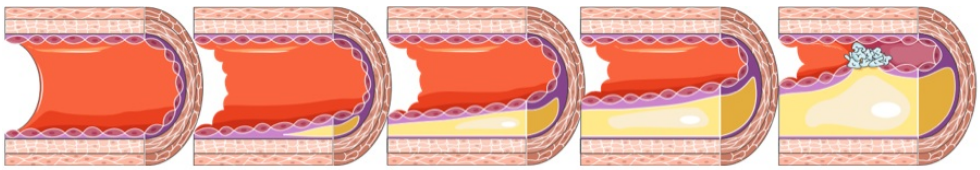


As aurículas têm como função receber o sangue que chega ao coração e os ventrículos a de bombear esse sangue do coração para os outros órgãos. O par aurícula/ventrículo direitos recebe o sangue de todo o corpo e envia-o aos pulmões para este ser oxigenado, enquanto o par aurícula/ventrículo esquerdos recebe o sangue oxigenado dos pulmões e envia-o para todo o corpo.

Tal como as restantes células do nosso organismo, o músculo cardíaco ou miocárdio também necessita de oxigénio e nutrientes para poder funcionar em pleno e manter todo o organismo oxigenado. No entanto, o coração não consegue extrair o oxigénio a partir das aurículas ou dos ventrículos e por isso possui um sistema próprio de vasos sanguíneos, que são as artérias coronárias.

As artérias coronárias sem doença têm paredes finas e extensíveis, o que facilita o fluxo sanguíneo. Com o passar dos anos, a parede destas artérias torna-se mais espessa por deposição de colesterol e outros componentes, formando as chamadas placas de ateroma, num processo designado por aterosclerose.

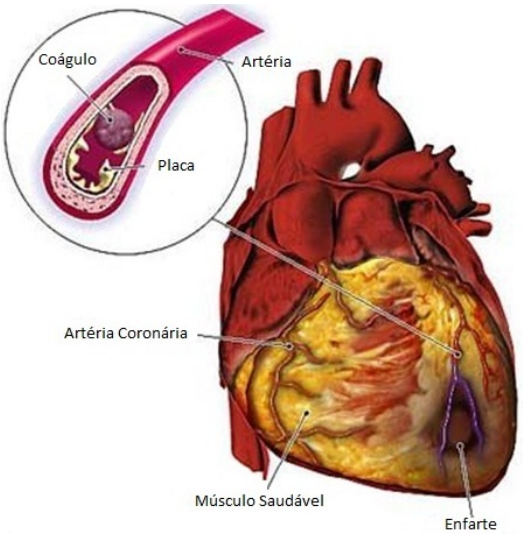
São vários os fatores de risco que levam ao espessamento das artérias. Alguns não é possível alterá-los, tais como a idade, o género ou os antecedentes familiares; outros podem ser alterados ou controlados com medicação: tabagismo, hipertensão arterial, diabetes, dislipidémia, obesidade, sedentarismo ou *stress*.



A doença aterosclerótica pode afetar as artérias coronárias, as artérias cerebrais ou as dos membros inferiores. Quando são atingidas as artérias de circulação cerebral, fala-se em acidente vascular cerebral (AVC). Se a doença atingir os membros inferiores, a doença aterosclerótica designa-se por claudicação intermitente e no caso das artérias coronárias a doença designa-se por doença coronária.

Na doença coronária existem dois tipos de evolução: AI e EAM. As duas situações devem-se ao crescimento da placa nas paredes das artérias ou à sua rutura. Quando o crescimento da placa ocupa mais de 70 % do diâmetro da artéria, o fluxo de sangue diminui drasticamente e, em situações em que o coração precisa de oxigenar mais o organismo, o músculo cardíaco entra em sofrimento e surge a dor no peito ou angina de peito.

Mesmo quando as placas não ocupam mais de 70 % do diâmetro da artéria, existe o risco de se romperem, o que ativa a defesa do organismo como se de uma ferida na pele se tratasse. As plaquetas e os fatores de coagulação tentam cobrir a zona da placa que se rompeu formando coágulos que interrompem o fluxo cardíaco. Quando a oclusão dura pouco tempo e o músculo cardíaco recupera, tem-se uma AI. Mas quando esta situação se prolonga por mais de 30 minutos, a lesão cardíaca torna-se irreversível e surge o EAM. Este define-se como a morte do músculo cardíaco que ficou privado de oxigénio e nutrientes. Esta zona do músculo cardíaco perde a sua capacidade de contração para bombear o sangue e é substituído por uma cicatriz.



4.2.2 Diagnóstico do enfarte

A rapidez com que se trata um enfarte é determinante no prognóstico. Quanto maior for a rapidez no tratamento, menor será a área do músculo cardíaco que é afetada. Alguns EAMs são súbitos e acompanhados de dor intensa no peito, mas noutros os sintomas surgem de forma progressiva e a dor ou desconforto não são muito intensas, levando o doente a desvalorizar as suas queixas. Os sintomas mais frequentes são: desconforto ou dor torácica; desconforto noutras áreas do tronco como um ou ambos os braços, pescoço, mandíbula, entre as omoplatas ou imediatamente abaixo do esterno; falta de ar; outros sinais e sintomas como hipersudorese (suores frios), náuseas e vómitos. Podem ainda surgir arritmias que conduzem a paragem cardíaca.

O diagnóstico de EAM é feito com base nos sintomas e em exames complementares como o eletrocardiograma (ECG) e análises sanguíneas para determinar os marcadores de morte miocárdica (componentes do músculo cardíaco que são libertados para a circulação quando há lesão das células cardíacas). Estes exames ajudam a determinar se se trata efetivamente de um EAM e, caso este diagnóstico se confirme, qual o tipo de abordagem mais adequado.

Outro exame importante para a doença coronária é o cateterismo cardíaco. Este é um exame invasivo que usa os raios X para avaliar o coração e as artérias coronárias, podendo no caso do EAM ajudar a determinar a localização da obstrução.

4.2.3 Tratamento e medicação

O tratamento inicial passa pelo restabelecimento precoce do fluxo sanguíneo. Existem duas formas possíveis de remoção do coágulo: a trombólise e a angioplastia coronária (PTCA). Em situações mais graves recorre-se à cirurgia de revascularização miocárdica, também designada por cirurgia de *bypass* (CABG) das artérias coronárias.

A doença coronária, uma vez estabelecida, é um processo crónico que exige uma vigilância durante toda a vida. Após um EAM, uma PTCA ou uma CABG, é fundamental a adesão a um plano terapêutico, que visa uma redução do risco de novos eventos cardiovasculares e uma melhoria da condição de saúde geral do doente. Existem diversas classes de medicamentos com efeitos distintos no organismo e muitas vezes complementares, daí que seja de extrema importância o cumprimento rigoroso dessa medicação. A terapêutica instituída nos doentes pode incluir: antiplaquetários, beta-bloqueantes, estatinas, inibidores da enzima da conversão da angiotensina e antihipertensores.

Atualmente o tempo médio de internamento para um enfarte não complicado é de 5 a 7 dias. Este período é importante porque é durante esta fase que mais frequentemente ocorrem as complicações cujo tratamento eficaz apenas é conseguido em ambiente hospitalar. As complicações mais frequentes são as arritmias, a insuficiência cardíaca, hemorragia, reenfarte, choque cardiogénico ou morte.

4.3 Descrição do estudo

Na prática clínica diária do serviço de cardiologia dos Hospitais da Universidade de Coimbra, os doentes admitidos no serviço com diagnóstico de SCA são incluídos continuamente entre 1 de janeiro de 2004 e 1 de setembro de 2006. Neste estudo foram registadas todas as características observadas, tais como antecedentes cardiovasculares, fatores de risco, características demográficas, eletrocardiograma, exame físico, análises laboratoriais, medicação e procedimentos invasivos e não invasivos implementados no tratamento do doente. Trata-se por isso de um estudo observacional. Após a alta hospitalar o seguimento dos doentes foi feito por telefone com uma periodicidade de seis a doze meses. Neste seguimento foram registados dados referentes à evolução da doença, como por exemplo, manutenção da medicação, ocorrência de complicações e valores laboratoriais para vigilância dos fatores de risco. Trata-se por isso de um estudo longitudinal prospetivo. Foram várias as fontes de recolha de informação posterior à alta: outro hospital, médico de família, médico de clínica privada ou o próprio doente. Os doentes considerados neste estudo tiveram data de admissão compreendida entre 1 de janeiro de 2004 e 31 de dezembro de 2005, tendo sido estabelecido o último contacto telefónico em dezembro de 2010.

4.3.1 Critérios de inclusão e exclusão

Foram incluídos todos os doentes com idade superior a 17 anos, com diagnóstico de SCA com menos de 48 horas de evolução desde o início dos sintomas. O diagnóstico de SCA foi definido como a presença de angina de peito em repouso nas últimas 48 horas associada a alterações no ECG caracterizadas por desvios do segmento ST ou ondas T negativas e/ou elevação de um biomarcador cardíaco. Considera-se diagnóstico de SCA com supra de ST quando existe elevação persistente do segmento ST (superior a 30 minutos), caso contrário são considerados SCA sem supra de ST (AIs ou EAMs sem supra de ST). Na ausência de angina de peito, considerou-se como SCA a elevação consistente da troponina cardíaca acima do valor de referência ou da CK-MB acima de duas vezes o valor de referência, associada a outras manifestações clínicas como desconforto torácico mal definido ou dispneia.

Excluíram-se os doentes com EAM após PTCA ou CABG; ou com EAM secundário provocado pelo aumento da necessidade de oxigénio ou pela diminuição no seu fornecimento ao organismo provocados, por exemplo, por espasmo de uma artéria coronária, níveis de hemoglobina baixos, arritmia, tensão arterial alta ou tensão arterial baixa.

4.3.2 Informação recolhida

Neste trabalho apenas foram consideradas as variáveis observadas durante a hospitalização do doente, excluindo-se a medicação, o tratamento e os resultados laboratoriais registados no internamento.

Variáveis demográficas

Género	É uma variável dicotómica (0=Masculino; 1=Feminino). As mulheres apresentam mais fatores de risco, uma manifestação da doença coronária mais tardia e sintomas de EAM diferentes dos homens, o que por vezes atrasa a resposta médica.
Idade à data de admissão	É uma variável contínua, medida em anos. Pode também define-se como variável categórica ordinal com categorias: < 65; $[65; 75[$ e ≥ 75. É um fator de risco importante por acelerar a aterosclerose, principalmente em homens com mais de 45 anos e mulheres com mais de 55 anos ou após a menopausa.
Índice de Massa Corporal	É medido em kg/m^2 e calculado pela fórmula peso(kg)/[altura(m)]^2. É uma variável contínua, que pode tomar valores entre 16 e 60. Pode ser definida como variável categórica com as seguintes categorias: < 25.0 (peso normal); $[25.0;30.0[$ (excesso de peso) e ≥ 30 (obesidade). O excesso de peso e a obesidade são responsáveis pelo aparecimento de alguns fatores de risco da aterosclerose.

Antecedentes

A existência destes antecedentes confere ao indivíduo, do ponto de vista clínico, um risco aumentado

de vir a desenvolver uma doença cardiovascular no futuro, podendo uma pessoa apresentar mais de um em simultâneo. Existem antecedentes cardiovasculares e não cardiovasculares. Todos os antecedentes considerados são variáveis dicotómicas.

Antecedentes não cardiovasculares	*Dislipidémia*; *Diabetes* (do tipo 1 ou tipo 2); *Tabagismo* (fumador ativo); *Antecedentes familiares*; *Stress*; *Hipertensão (HTA).*
Antecedentes cardiovasculares	*AVC/AIT*; *Doença arterial periférica (DAP)*; *Insuficiência cardíaca (ICC).*

Não se consideraram outros antecedentes cardiovasculares porque se assume que não existem antecedentes de EAM, de cateterismo, de PTCA nem de CABG.

Variáveis obtidas no exame físico

O exame físico é um conjunto de técnicas utilizadas por profissionais de saúde no diagnóstico de uma doença. Tem como objetivo a deteção de anomalias, de modo a decidir quais as intervenções mais adequadas para o tratamento e prevenção do agravamento do estado de saúde do paciente. As variáveis seguintes avaliam-se na admissão hospitalar.

Classe kk	É uma escala que avalia a gravidade da insuficiência cardíaca, segundo quatro valores possíveis: I, II, III e IV. Existe insuficiência cardíaca quando a classe kk é II, III ou IV.
TA sistólica	É medida em mmHg. Trata-se de uma variável contínua que pode tomar valores entre 50 e 280. Também é usual definirem-se as seguintes categorias: < 120; $[120; 140[$ e ≥ 140.
TA diastólica	É avaliada em simultâneo com a TA sistólica e também é medida em mmHg. Trata-se de uma variável contínua que pode tomar valores entre 25 e 140. Podem definir-se as categorias: < 70; $[70; 90[$ e ≥ 90.
Frequência cardíaca	É medida em batimentos por minuto (bpm). É uma variável contínua que pode tomar valores entre 10 e 200. Como variável categórica podem considerar-se as seguintes categorias: < 65; $[65; 85[$ e ≥ 85.

Variável diagnóstico

Os doentes são classificados segundo um de três diagnósticos possíveis.

Diagnóstico na admissão	EAM com supra de ST; EAM sem supra de ST; Angina instável.

Variáveis laboratoriais

Durante a admissão hospitalar do doente são realizadas algumas análises clínicas que permitem fazer uma avaliação do estado de saúde geral do doente.

Creatinina	Permite avaliar a função renal. É medida em mg/dL e pode tomar valores entre 0.4 e 20.0. As categorias frequentemente consideradas são: < 1.0; $[1.0; 1.4[$ e ≥ 1.4. Os valores normais de creatinina situam-se entre 0.6 e 1.1 mg/dL nos homens e entre 0.5 a 0.9 mg/dL nas mulheres.
TFG	Tal como a creatinina, avalia a função renal. Mede-se em mL/min/m^2. Pode tomar valores entre 3 e 150. Podem ser definidas as seguintes categorias: < 50; $[50; 70[$ e ≥ 70. Quando mais baixo o valor de TFG, pior é a função renal.
Glicémia	Trata-se de uma variável contínua que pode tomar valores entre 20 e 1000. Também é usual definir-se como variável categórica com as seguintes categorias: < 110; $[110; 126[$ e ≥ 126. A presença de glicémia aumentada ou de diabetes agrava o risco de doenças cardiovasculares.
Colesterol total	É o principal parâmetro clínico de avaliação do perfil lipídico. A unidade de medida é mg/dL e pode tomar valores entre 50 e 500. Como variável categórica consideram-se as seguintes categorias: < 190; $[190; 240[$ e ≥ 240. O risco de aterosclerose aumenta com o aumento de CT. A avaliação de CT, juntamente com o HDL, o LDL e os triglicerídeos, caracterizam o perfil lipídico.
Colesterol HDL	É uma variável contínua, medida em mg/dL. Pode tomar valores entre 10 e 120. Usualmente consideram-se as seguintes categorias: ≤ 45; $]45; 55]$ e > 55. O risco de aterosclerose aumenta com a diminuição de HDL.
Colesterol LDL	Toma valores entre 20 e 300 e a sua unidade de medida é mg/dL. Usualmente consideram-se as seguintes categorias: < 100; $[100; 130[$ e ≥ 130. O risco de aterosclerose aumenta com o aumento de LDL.
Triglicerídeos	Pode tomar valores entre 20 e 2000, medidos em mg/dL. As categorias mais usadas são: < 150; $[150; 250[$ e ≥ 250. Valores elevados estão associados a um maior risco de doença aterosclerótica.

Tempo de vida

Neste trabalho tem-se como principal objetivo a determinação dos fatores que têm influência significativa no tempo até à ocorrência de enfartes múltiplos, medido desde a data de admissão hospitalar. Assim, além da observação de todas as variáveis referidas anteriormente, registaram-se todos os enfartes decorrentes da SCA que originou o internamento hospitalar. Foram registadas as datas em que foram estabelecidos contactos com cada doente durante a fase de seguimento e as datas de todos os

enfartes ocorridos. Na amostra considerada foram observados até três enfartes para o mesmo doente. Assim, para cada enfarte, registaram-se os instantes em que os doentes entraram no conjunto de indivíduos em risco e em que ocorreram os enfartes e determinou-se o correspondente número de dias em que estiveram em risco. Depois do último enfarte observado, registou-se também o intervalo entre este último enfarte (independentemente de ser o primeiro, segundo ou terceiro) e o último contacto com o doente e o número de dias correspondente.

Para os doentes que não sofreram qualquer enfarte, considerou-se apenas um intervalo de tempo e determinou-se o correspondente número de dias decorridos desde a admissão hospitalar até ao último contacto com o doente.

4.4 Resultados

4.4.1 Caracterização da amostra

Foram admitidos 895 doentes com diagnóstico de SCA no serviço de cardiologia dos Hospitais da Universidade de Coimbra, entre 1 de janeiro de 2004 e 1 de setembro de 2006. Destes, 813 (90.84%) não sofreram enfartes durante o período de seguimento, 68 (7.60%) sofreram um enfarte, 11 (1.23%) sofreram dois enfartes e 3 (0.33%) sofreram três enfartes. A duração mediana do internamento foi de 5 dias, enquanto o seguimento clínico mediano, após a admissão hospitalar, foi de 601 dias. O último contacto foi realizado a 1 de setembro de 2007.

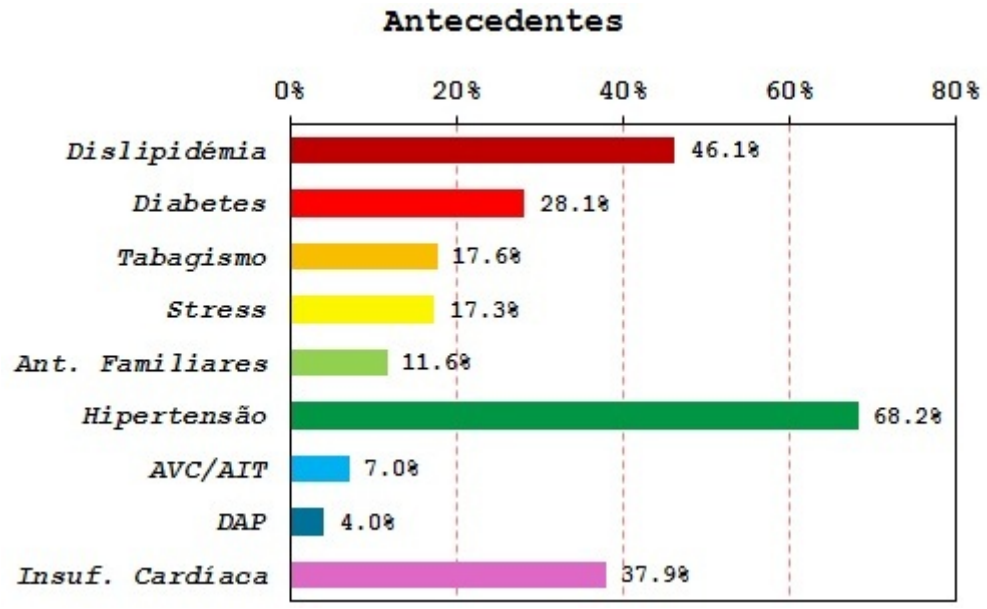
Das covariáveis observadas sete têm valores omissos, como mostra a tabela seguinte:

Creatinina	76 (8.5%)
TFG	76 (8.5%)
Glicémia	75 (8.4%)
CT	147 (16.4%)
HDL	148 (16.5%)
LDL	148 (16.5%)
TG	147 (16.4%)

Dos 895 doentes hospitalizados, 317 (38.2%) apresentavam diagnóstico de EAM com supra de ST, 411 (45.9%) foram diagnosticados com EAM sem supra de ST e os restantes 167 (18.7%) com AI.

Nesta amostra 32.1% são mulheres, enquanto 67.9% são homens. Os antecedentes mais comuns são a hipertensão, a dislipidémia e a insuficiência cardíaca e os menos frequentes são a DAP e o AVC/AIT.

No gráfico seguinte está representada a caracterização de todos os antecedentes observados.



Na tabela seguinte encontra-se a caracterização da idade e do índice de massa corporal.

	Idade	*IMC*
Média	67	27.4
Desvio Padrão	12	4.2
P_{25}	59	24.6
Mediana	68	27.2
P_{75}	76	29.4
Mínimo	27	17.0
Máximo	93	64.1

Segundo o indicador de insuficiência cardíaca na admissão hospitalar, a distribuição dos doentes pelas quatro classes kk é a seguinte: 770 (86.0%) apresentaram classe I, 108 (12.1%) classe II, 10 (1.1%) classe III e 7 (0.8%) classe IV. Foi observada insuficiência cardíaca na admissão em 125 (14.0%) doentes. Relativamente à tensão arterial e à frequência cardíaca, as respetivas caracterizações encontram-se na tabela seguinte:

	TAS	*TAD*	*FC*
Média	138	74	78
Desvio Padrão	24	14	15
P_{25}	120	64	68
Mediana	135	72	77
P_{75}	151	80	86
Mínimo	48	30	35
Máximo	223	130	140

A função renal e a glicémia têm a seguinte caracterização:

	Creatinina	*TFG*	*Glicémia*
Média	1.2	59.8	138
Desvio Padrão	0.8	21.5	68
P_{25}	0.9	46.6	101
Mediana	1.0	63.1	119
P_{75}	1.2	74.4	153
Mínimo	0.5	4.1	66
Máximo	10.4	111.6	998

Cerca de 25% dos doentes têm creatinina superior a 1.2 mg/dL e 25% têm valor inferior a 0.9 mg/dL. Por outro lado, 25% têm TFG inferior a 46.6 mL/min/m^2. A glicémia no sangue é, em 50% dos doentes, inferior a 119 mg/dL. Quanto ao perfil lipídico, a sua caracterização encontra-se na tabela seguinte:

	CT	*HDL*	*LDL*	*TG*
Média	192	43	129	173
Desvio Padrão	51	10	37	134
P_{25}	156	35	102	102
Mediana	186	42	126	139
P_{75}	220	48	151	197
Mínimo	88	12	33	37
Máximo	612	91	291	1767

Nesta amostra, cerca de 25% dos doentes têm valor de colesterol total inferior a 156 mg/dL e aproximadamente metade tem valor acima do normal, que é 190 mg/dL. Em 25% dos doentes foi observado valor de HDL inferior a 35 mg/dL, que é bastante inferior a 45 mg/dL, que é o valor acima do qual o HDL é normal. Relativamente ao valor de LDL, aproximadamente 75% dos doentes tem valor observado normal (150 mg/dL). Cerca de metade dos doentes tem valor de triglicerídeos compreendido entre 100 mg/dL e 200 mg/dL.

Os histogramas das variáveis contínuas podem ser consultados no apêndice D.1.

4.4.2 Modelo de Cox para o tempo até à ocorrência do primeiro enfarte

A relação entre cada uma das variáveis explanatórias e o tempo até ao primeiro enfarte foi modelada através de um modelo de Cox univariável. Dos 895 doentes, 813 (90.84%) não sofreram enfartes durante o período de seguimento e 82 (9.16%) sofreram um enfarte. Como, nesta análise, se considerou o tempo até ao primeiro enfarte, a mediana do tempo de seguimento passou a ser 569 dias.

Análise univariável (1.º Modelo de Cox)

Na tabela seguinte encontram-se os resultados obtidos pelo ajustamento dos vários modelos de Cox aos dados:

Covariáveis	*Valor p*	*RR (IC95%)*
Idade	<0.001	1.044 (1.023;1.065)
IMC	0.692	1.011 (0.959;1.065)
Género	0.959	0.988 (0.619;1.578)
Dislipidémia	0.088	0.684 (0.442;1.058)
Diabetes	0.005	1.872 (1.211;2.894)
Tabagismo	0.632	0.873 (0.500;1.524)
Stress	0.176	0.646 (0.343;1.217)
Ant. Familiares	0.099	0.497 (0.217;1.140)
Hipertensão	0.457	1.196 (0.746;1.917)
AVC/AIT	0.919	1.049 (0.420;2.616)
DAP	0.369	1.588 (0.579;4.354)
Insuf. Cardíaca	0.620	1.121 (0.713;1.763)
TA Sistólica	0.557	0.997 (0.988;1.006)
TA Diastólica	0.159	0.989 (0.974;1.004)
Freq. Cardíaca	0.171	0.990 (0.976;1.004)
Classe KK II	0.212	1.460 (0.806;2.645)
Classe KK III	0.356	1.950 (0.472;8.056)
Classe KK IV	0.594	1.711 (0.237;12.342)
Classe KK I	Classe de referência	
Diag: EAM com supST	0.084	1.931 (0.915;4.074)
Diag: EAM sem supST	0.015	2.421 (1.186;4.944)
Diag: Angina Instável	Classe de referência	
Creatinina	0.154	1.183 (0.939;1.489)
TFG	0.017	0.988 (0.978;0.998)
Glicémia	0.008	1.002 (1.001;1.004)
Colesterol Total	0.409	0.998 (0.993;1.003)
Colesterol HDL	0.035	0.974 (0.951;0.998)
Colesterol LDL	0.766	0.999 (0.993;1.005)
Triglicerídeos	0.970	1.000 (0.998;1.002)

As variáveis que, isoladamente, têm influência significativa no tempo são: a idade, a dislipidémia, a diabetes, os antecedentes familiares, o diagnóstico, a creatinina, a TFG, a glicémia e o colesterol HDL. Quando se comparam os doentes com AI com os doentes de cada um dos outros diagnósticos em separado, verifica-se um aumento significativo do risco entre os EAM sem supra de ST. O risco estimado de

ocorrência de enfarte entre os doentes com EAM sem supra ST é 2.421 vezes o risco dos doentes com AI. O risco estimado de ocorrência de enfarte dos doentes com EAM com supra de ST é aproximadamente o dobro do dos doentes com AI. Por seu lado, os doentes diabéticos apresentam um acréscimo de 87.2% no risco de enfarte em relação aos não diabéticos. Estima-se que o aumento da idade está associado a um aumento do risco de enfarte. Quanto à TFG, os doentes com valores mais altos tendem a ter um menor risco de enfarte, o que significa que têm um melhor prognóstico. Os doentes com valores mais elevados de glicémia estão também associados a um aumento do risco de enfarte, enquanto os doentes com valores mais altos de HDL apresentam melhor prognóstico.

Análise multivariável (1.º Modelo de Cox)

Depois de estimar o efeito que cada covariável, por si só, tem no tempo até à ocorrência de enfarte, interessa agora construir um modelo de regressão de Cox multivariável. Após ser feita a seleção de variáveis segundo o método *stepwise forward* com a utilização do teste de razão de verosimilhanças, é obtido um modelo de Cox a que correspondem os seguintes resultados:

Covariáveis	$\hat{\beta}_j$	$EP(\hat{\beta}_j)$	*Valor p*	*RR (IC95%)*
Idade	0.036	0.011	0.001	1.037 (1.014;1.060)
Freq. Cardíaca	-0.017	0.008	0.033	0.983 (0.968;0.999)
Insuf. Cardíaca	0.490	0.238	0.039	1.632 (1.024;2.600)
Colesterol HDL	-0.024	0.012	0.048	0.976 (0.952;1.000)
Glicémia	0.002	0.001	0.010	1.002 (1.001;1.004)

Note-se que, quando consideradas em conjunto com outras covariáveis, a insuficiência cardíaca e a frequência cardíaca passam a ter influência significativa no tempo até enfarte. Antes de interpretar os resultados do modelo de Cox multivariável é necessário fazer a análise dos resíduos, para testar a proporcionalidade das funções de risco para todas as covariáveis, avaliar a forma funcional das covariáveis contínuas, identificar os valores extremos e as observações influentes.

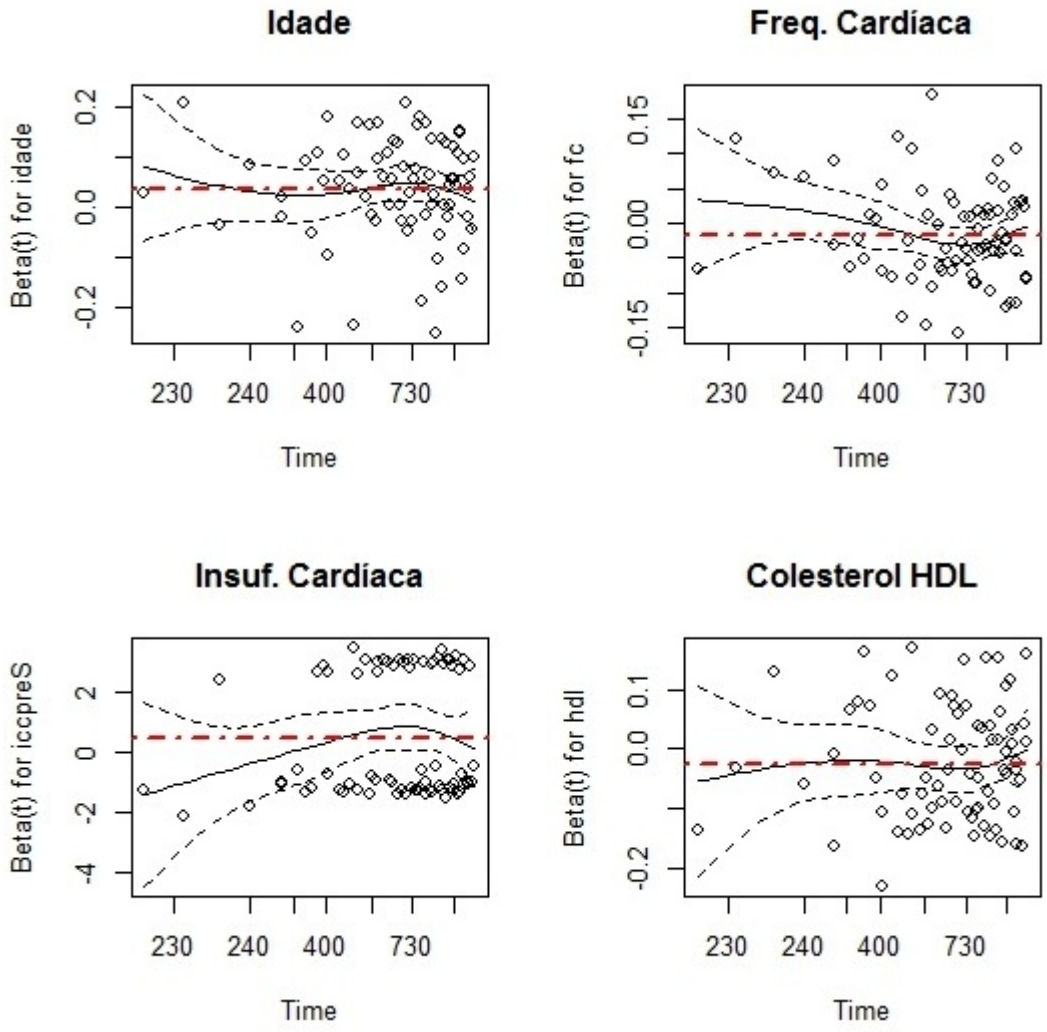
Resíduos de Schoenfeld ponderados (1.º Modelo de Cox)

O primeiro passo na análise dos resíduos refere-se à análise da proporcionalidade, que é testada na globalidade e individualmente para cada covariável. Neste último caso, o teste será acompanhado por um gráfico em que se representam os resíduos, os intervalos de confiança da curva de suavização *spline* dos resíduos e uma linha horizontal vermelha correspondente ao efeito constante da covariável estimado pelo modelo. Não existe violação da proporcionalidade quando a linha horizontal se mantém dentro do intervalo de confiança em todos os instantes. A verificação da hipótese de proporcionalidade de cada uma das covariáveis deve ser feita considerando tanto o teste como o gráfico correspondente. Os

resultados do teste da hipótese de proporcionalidade são:

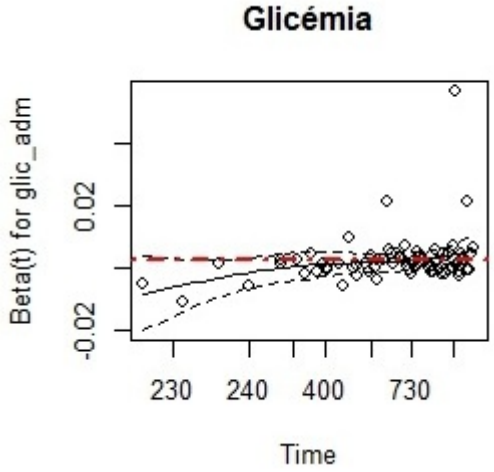
	rho	*chisq*	*Valor p*
Idade	-0.031	0.077	0.781
Freq. Cardíaca	-0.155	1.760	0.185
Insuf. Cardíaca	0.131	1.257	0.262
Colesterol HDL	0.046	0.132	0.717
Glicémia	0.277	5.401	0.020
GLOBAL	NA	7.099	0.213

Nesta tabela, a coluna "rho" contém o coeficiente de correlação linear entre os resíduos e o tempo de vida e as colunas "chisq" e "Valor p" contêm o valor observado da estatística qui-quadrado e o respetivo valor p. Constata-se que na globalidade o modelo não viola a hipótese de proporcionalidade das funções de risco. Em seguida apresentam-se os gráficos dos resíduos de Schoenfeld.



A partir da análise dos gráficos, conclui-se que os efeitos da idade e do colesterol HDL são constantes ao longo do tempo. Em relação à frequência cardíaca e à insuficiência cardíaca, o efeito parece ser

diferente no início da observação, no entanto, como existem poucas observações, não se valorizam estas variações. Além disso, os testes acima considerados não rejeitam a hipótese de proporcionalidade dos riscos. Em relação à glicémia, o teste leva à rejeição da hipótese de proporcionalidade; no entanto, a partir do gráfico seguinte,

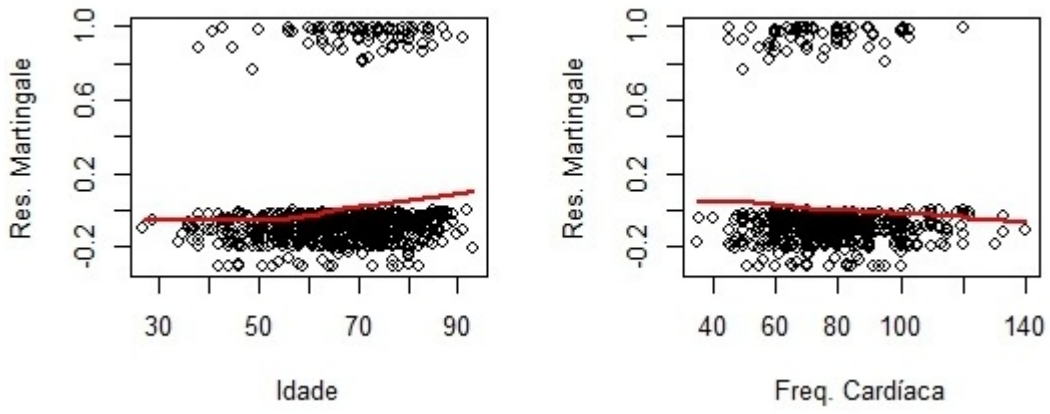


é possível constatar que nos instantes iniciais a variação observada no efeito da covariável se deve ao facto de existirem poucos valores e, portanto, essa variação pode ser atribuída à flutuação aleatória. Além disso, o teste global de proporcionalidade não rejeita esta hipótese, pelo que esta pequena variação da glicémia pode ser desvalorizada.

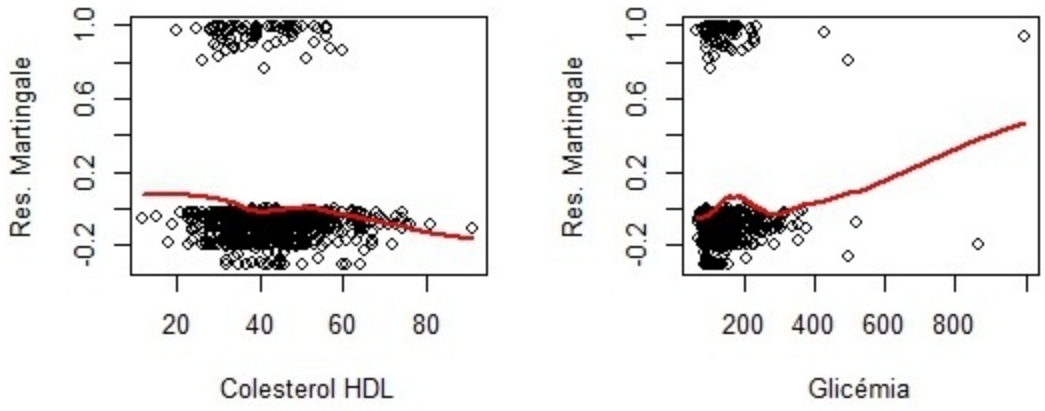
Resíduos martingala (1.º Modelo de Cox)

Os resíduos martingala do modelo nulo permitem explorar a forma funcional de cada uma das covariáveis contínuas, avaliando a forma como a covariável se associa ao tempo. Permitem assim concluir se é necessário fazer alguma transformação usando, por exemplo, a raiz quadrada ou o logaritmo. Os gráficos para esta avaliação colocam no eixo das abcissas os valores da covariável cuja forma se pretende analisar e nas ordenadas os resíduos martingala do modelo nulo. Para facilitar a análise destes resíduos, representa-se também uma curva *lowess*.

De seguida são apresentados os gráficos dos resíduos martingala, um por cada covariável contínua.



A partir dos gráficos das covariáveis idade e frequência cardíaca, não existem dúvidas que a forma funcional é linear, não sendo necessário proceder a uma transformação.



Relativamente ao colesterol HDL, a curva é monótona decrescente até valores próximos de 40 mg/dL, valor a partir do qual a função se mantém paralela ao eixo das abcissas até atingir aproximadamente 55 mg/dL de HDL e começar novamente a decrescer ligeiramente. Apesar destas flutuações, o aspeto global desta função leva a concluir que não é necessário transformar a covariável HDL. Finalmente, em relação à glicémia, a curva apresenta uma forma marcadamente não linear e por esse motivo esta variável deve ser transformada. De entre as várias transformações consideradas optou-se pelo inverso da glicémia por ser a mais adequada.

Análise multivariável (2.º Modelo de Cox)

Construiu-se novamente o modelo de regressão de Cox multivariável, usando o método de seleção de variáveis *stepwise forward*, mas desta vez considerou-se a covariável resultante da transformação da glicémia em vez da covariável glicémia original. O modelo final é:

Covariáveis	$\hat{\beta}_j$	$EP(\hat{\beta}_j)$	*Valor p*	*RR (IC95%)*
Idade	0.035	0.011	0.002	1.035 (1.014;1.060)
Freq. Cardíaca	-0.019	0.008	0.016	0.981 (0.968;0.999)
Insuf. Cardíaca	0.510	0.239	0.033	1.666 (1.024;2.600)
Colesterol HDL	-0.024	0.012	0.052	0.977 (0.952;1.000)
1/Glicémia	-146.700	50.560	0.004	(*)

(*) *RR(IC95%)*: 1.897×10^{-64} $(1.728\times10^{-107};2.083\times10^{-21})$.

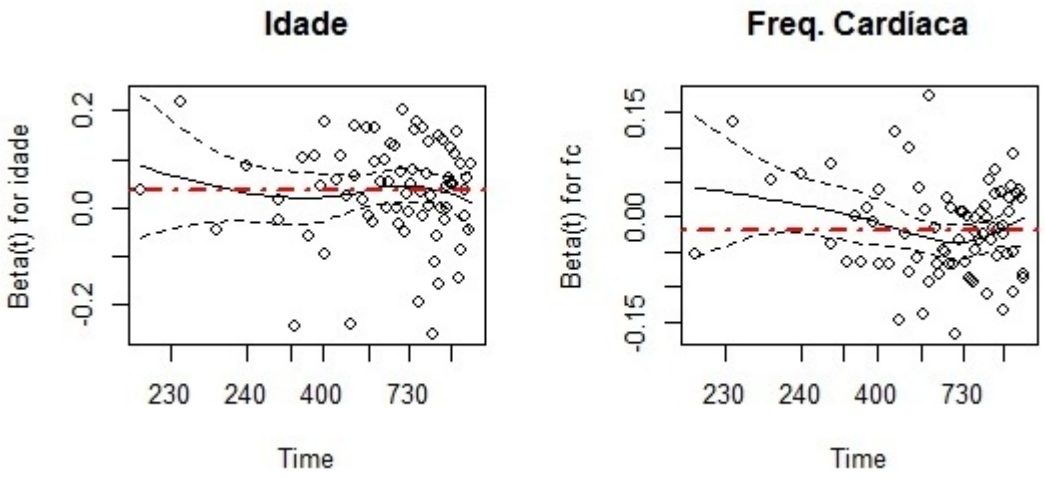
As covariáveis incluídas neste modelo são as mesmas que haviam sido incluídas no primeiro modelo de Cox ajustado, tendo apenas sido substituída a glicémia pela sua transformada. Apesar disso é necessário repetir a análise dos resíduos de Schoenfeld e martingala para todas as variáveis.

Resíduos de Schoenfeld ponderados (2.º Modelo de Cox)

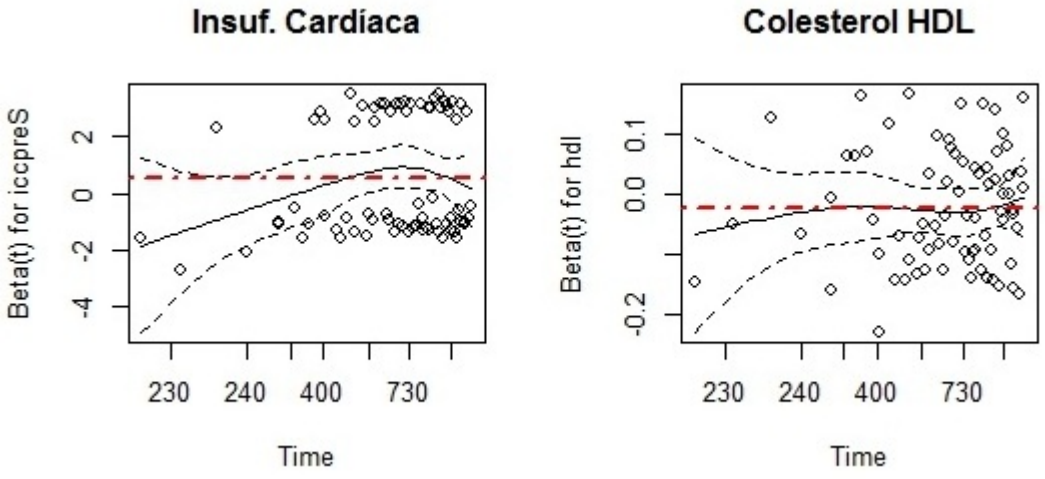
Os resultados do teste da hipótese de proporcionalidade são:

	rho	*chisq*	*Valor p*
Idade	-0.031	0.081	0.776
Freq. Cardíaca	-0.155	1.722	0.189
Insuf. Cardíaca	0.167	2.084	0.149
Colesterol HDL	0.061	0.224	0.636
1/Glicémia	-0.120	1.060	0.303
GLOBAL	NA	4.218	0.518

Constata-se que na globalidade o modelo não viola a hipótese de proporcionalidade das funções de risco. Veja-se os gráficos dos resíduos de Schoenfeld para as seguintes covariáveis:

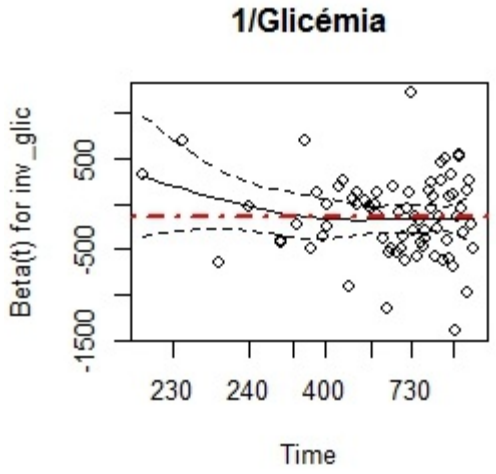


A partir dos dois gráficos conclui-se que as duas variáveis têm efeitos constantes ao longo do tempo. Apesar de parecer que o efeito da frequência cardíaca é diferente no início da observação, essa variação não é valorizada, o que também é confirmado pelo teste de proporcionalidade.



Em relação à insuficiência cardíaca e ao HDL também se conclui que os riscos são proporcionais. As variações observadas para a covariável HDL, à semelhança do que aconteceu com a frequência cardíaca, também não são consideradas significativas.

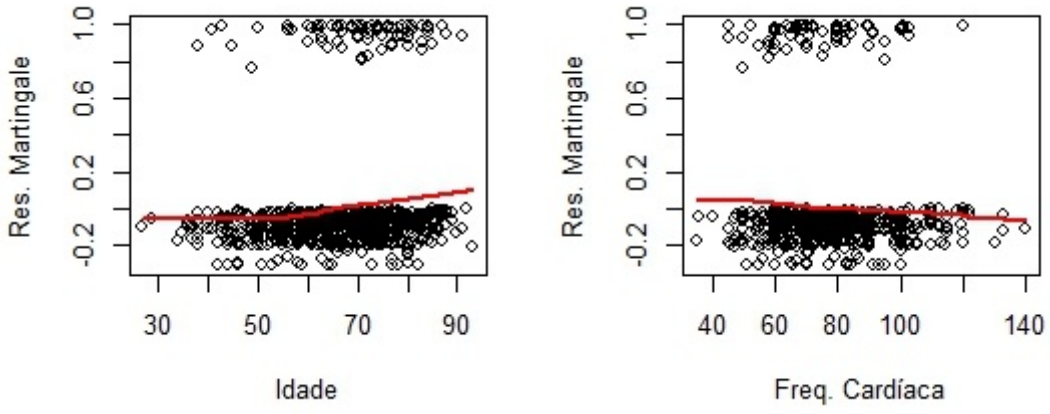
Finalmente, em relação à covariável resultante da transformação da glicémia, não existe evidência para a rejeição da proporcionalidade dos riscos. Ao observar o respetivo gráfico dos resíduos de Schoenfeld ponderados,



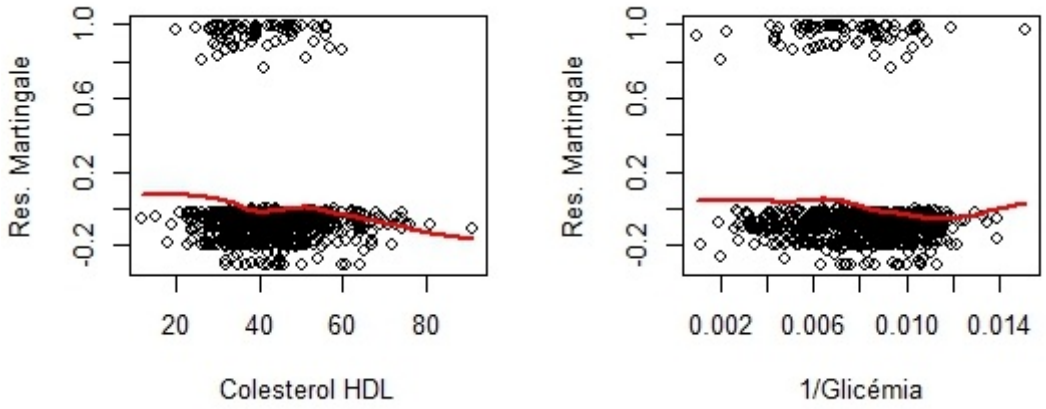
conclui-se que, de facto, o efeito desta covariável é constante.

Resíduos martingala (2.º Modelo de Cox)

Os gráficos dos resíduos martingala para cada covariável contínua são os seguintes:



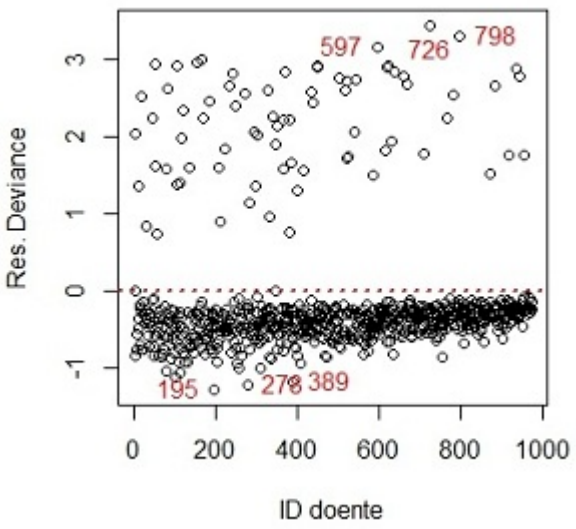
A partir dos gráficos das covariáveis idade e frequência cardíaca, não existem dúvidas que a forma funcional é linear, continuando a não ser necessário proceder a uma transformação. Em relação ao colesterol HDL e à transformada da glicémia, os seguintes gráficos permitem indagar sobre a necessidade de proceder a uma transformação da covariável HDL e confirmar se a transformação da glicémia foi adequada.



O gráfico do HDL é idêntico ao que foi obtido para o primeiro modelo de Cox ajustado, logo não é necessário transformar a covariável. Quanto à transformada da glicémia, observam-se algumas oscilações, sendo a mais marcada a partir do valor 0.011. No entanto, não deve ser valorizada por existirem poucas observações acima desse valor e existir um valor superior a 0.014 com resíduo martingala próximo de 1. Assim, considera-se que esta transformação da glicémia é adequada.

Resíduos *deviance* (2.º Modelo de Cox)

A partir dos resíduos *deviance* é possível identificar *outliers* e, portanto, identificar os doentes que estão mal ajustados pelo modelo de Cox. Os *outliers* com resíduo negativo correspondem, neste contexto, a indivíduos que sofreram um enfarte antes do que estava previsto, enquanto os que estão associados a um resíduo positivo, possuem características que reduziria o tempo até enfarte. No entanto, sofreram um enfarte tardiamente ou não chegaram a sofrer nenhum durante o tempo em que estiveram sob observação. Assim, a identificação dos indivíduos mal ajustados é feita a partir de um gráfico dos resíduos *deviance versus* o índice da observação.



A partir do gráfico é possível ver que os resíduos *deviance* não têm uma distribuição simétrica em torno de zero e isso deve-se à presença de uma grande percentagem de observações censuradas.
Ao consultar o gráfico destacam-se os doentes número 195, 278 e 889 com resíduos negativos e os doentes número 597, 726 e 798 com resíduos positivos. Na tabela seguinte apresentam-se, para estes seis doentes, os respetivos valores observados para as covariáveis incluídas no modelo multivariável.

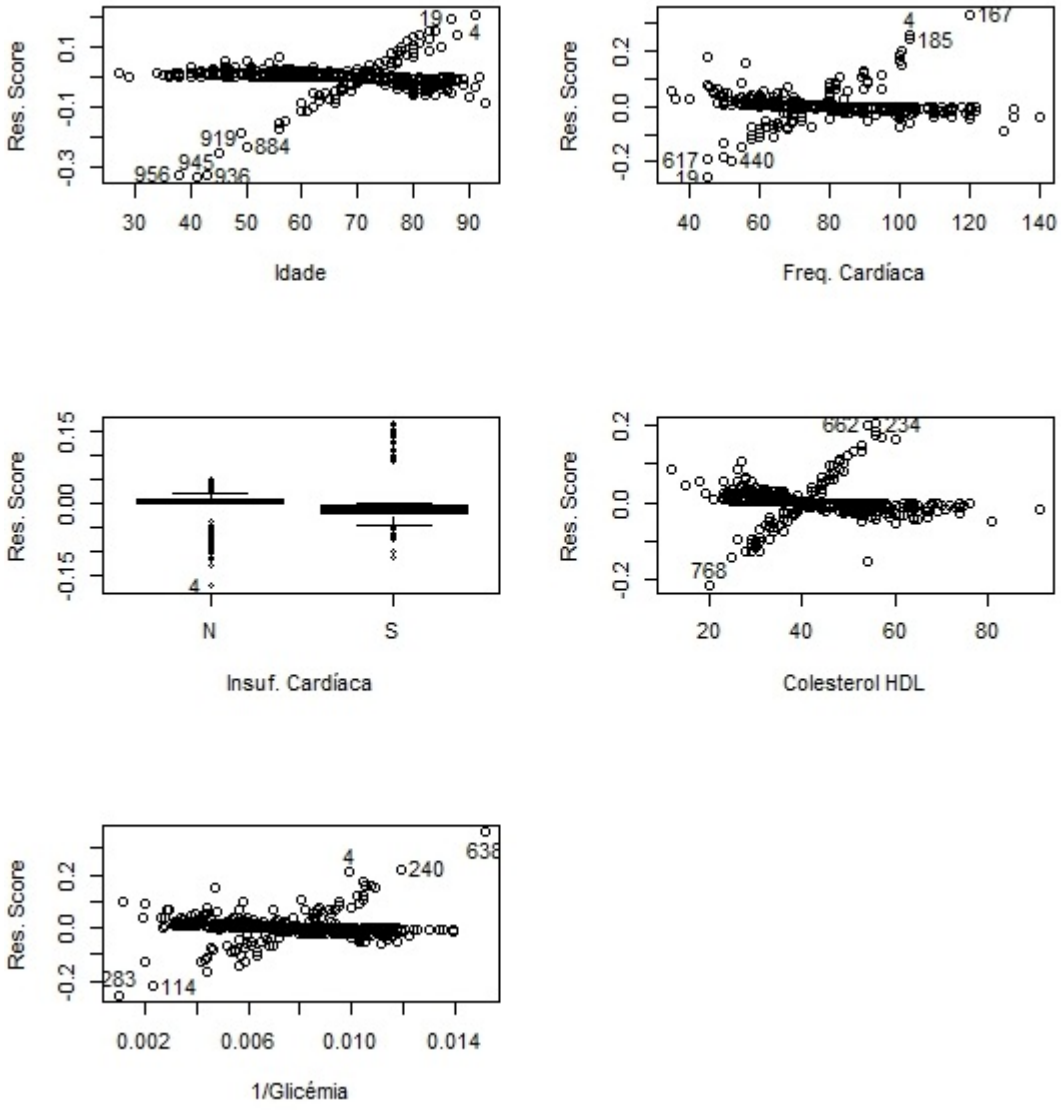
ID doente	*Idade*	*FC*	*ICC*	*HDL*	*Glic*	*Enfarte*	*Dias*
195	77	56	Sim	54	210	Não	825
278	74	68	Sim	27	177	Não	846
389	70	45	Sim	36	173	Não	465
597	65	69	Sim	49	147	Sim	9
798	56	80	Sim	47	139	Sim	6
726	60	90	Não	44	140	Sim	5

Nesta tabela, cada doente é identificado com um número que se encontra na coluna "ID doente". O resíduo obtido está registado na coluna "Res". Os valores que foram observados para as covariáveis idade, frequência cardíaca, insuficiência cardíaca, colesterol HDL e glicémia, encontram-se nas colunas "Idade", "FC", "ICC", "HDL" e "Glic", respetivamente. A coluna "Enfarte" indica se o doente sofreu ou não um enfarte, enquanto a coluna "Dias" contém o tempo de vida ou tempo de censura, em dias. Conclui-se, da tabela anterior, que os três doentes com resíduos positivos sofreram enfarte nos primeiros 10 dias após a admissão hospitalar. Mas isso não seria de esperar, dado que têm idades inferiores ou iguais a 65 anos, têm valores de glicémia próximos de 145 (superiores ao valor 126, a partir do qual se diagnostica a diabetes) e, não têm valores de HDL abaixo de 40. Em relação ao valor de FC, apenas o doente 597 tem valor abaixo de 70 bpm. Apenas o doente 726 não tem insuficiência cardíaca.
Quanto aos doentes com resíduos negativos, todos têm insuficiência cardíaca anterior à admissão hospitalar, têm idades superiores a 70 anos e glicémia superior a 170. As frequências cardíacas também são baixas, principalmente nos indivíduos 195 e 389. O valor de HDL é bastante baixo nos doentes 278 e 389. Perante os valores observados seria de esperar que tivessem sofrido um enfarte num tempo inferior ao observado. É de notar que se optou por considerar os resíduos *deviance* para identificar *outliers* em vez dos resíduos martingala. Isso deve-se ao facto do resíduo martingala tomar valores inferiores a 1, o que pode dificultar a identificação, no gráfico destes resíduos *versus* o índice da observação, dos *outliers* com resíduos positivos.

Resíduos *score* ponderados (2.º Modelo de Cox)

Os resíduos *score* desempenham um papel fundamental na verificação da influência que cada observação tem no ajustamento do modelo de Cox. Esta influência é analisada para cada parâmetro em separado.

Quanto mais próximo de zero estiver o resíduo *score*, menos influente será a observação na estimação do parâmetro correspondente. Assim, constroem-se tantos gráficos quantas as covariáveis incluídas no modelo, colocando-se os resíduos nas ordenadas e os valores de cada covariável no eixo das abcissas. Quando as covariáveis são contínuas, usam-se gráficos de dispersão e quando são categóricas usa-se uma *boxplot* para cada categoria. Na visualização gráfica, recorre-se frequentemente aos resíduos *score* ponderados por tornar mais fácil a identificação dos pontos influentes. Seguem-se abaixo os gráficos.



Os resíduos correspondentes à insuficiência cardíaca não são, em valor absoluto, muito elevados e por isso apenas se assinalou o doente 4, que não teve insuficiência cardíaca e sofreu um enfarte ao fim de 39 dias. Este doente é também considerado influente na obtenção das estimativas dos parâmetros associados à idade e à frequência cardíaca. Apresenta-se em seguida a análise dos pontos influentes para as restantes covariáveis.

Os doentes que têm maior influência na estimação do efeito da idade são:

ID doente	4	19	884	919	936	945	956
Idade	91	87	50	45	43	41	38
Enfarte	Sim	Sim	Sim	Sim	Sim	Sim	Sim
Dias	39	7	66	476	32	37	491

No caso da idade, foram mais influentes as observações de doentes que sofreram enfarte poucos dias depois da admissão hospitalar e que tinham idade inferior ou igual a 50 anos ou superior a 80 anos. Os doentes 919 e 956 têm tempos até enfarte mais longos, mas a sua idade é baixa, daí serem influentes. Nenhum destes doentes foi mal ajustado pelo modelo.

ID doente	4	19	167	185	440	617
Freq. Cardíaca	103	45	120	103	52	45
Enfarte	Sim	Sim	Sim	Sim	Sim	Sim
Dias	39	7	13	18	17	240

No caso da frequência cardíaca, o raciocínio é o mesmo. Existem três doentes com valores de FC muito baixos (considerando-se valores baixos os que são inferiores ou estão próximos de 50 bpm) e três com valores acima de 100 bpm. Estes doentes tendem a sofrer enfartes passados poucos dias da admissão hospitalar.

ID doente	234	662	768
HDL	56	54	20
Enfarte	Sim	Sim	Sim
Dias	48	15	55

No que diz respeito ao HDL os doentes influentes sofreram enfarte precocemente e têm valores próximos de 55 ou muito baixos, como é o caso do doente 768.

ID doente	114	240	283	638
Glicémia	427	84	998	66
Enfarte	Sim	Sim	Sim	Sim
Dias	207	40	278	82

Os doentes 240 e 638 são influentes porque têm valores de glicémia baixos e sofreram enfarte precocemente. Quanto aos doentes 114 e 283 a sua influência deve-se ao facto do valor da glicémia ser extremamente alto.

Interpretação dos resultados (2.º modelo de Cox)

Após a análise dos resíduos e depois de concluir que o modelo ajustado é adequado, está-se em condições de interpretar os resultados obtidos. Pode concluir-se que a idade, a frequência cardíaca, a presença de insuficiência cardíaca prévia, o colesterol HDL e a glicémia têm influência significativa no tempo até enfarte.

Os resultados que a seguir são indicados, para cada covariável, referem-se a indivíduos que têm valores iguais das restantes covariáveis.

- Ao compararmos dois indivíduos que diferem apenas num ano de idade, o mais velho tem um aumento estimado do risco de sofrer um enfarte de 3.5%. Quando se comparam indivíduos cuja diferença de idades é de 5 anos, o mais velho tem um acréscimo de 18.9% no risco de enfarte.
- No que diz respeito ao valor da frequência cardíaca na admissão, quando se comparam dois doentes com apenas 1 bpm de diferença, o que tem um valor mais baixo tem um aumento do risco de sofrer um enfarte de 1.9%. Quando a diferença é de 10 bpm, o acréscimo estimado do risco é de 17.1%.
- Estima-se que a presença de insuficiência cardíaca prévia em doentes com SCA aumenta o risco em 66.6%, quando comparado com os que não têm este antecedente.
- Em relação ao colesterol HDL, considerando dois doentes que difiram em 1 mg/dL, o que apresenta menor valor tem um acréscimo do risco de enfarte estimado em 2.3%. Quando a diferença observada é de 5 mg/dL, o aumento do risco é de 11.2%.
- A interpretação da influência da glicémia no tempo até enfarte depende dos valores considerados para os dois doentes que se pretende comparar. Quando diferem, em 40 mg/dL, o aumento estimado do risco é

    52.1%, quando os valores da glicémia são 100 mg/dL e 140 mg/dL;

    30.4%, quando os valores da glicémia são 130 mg/dL e 170 mg/dL;

    20.1%, quando os valores da glicémia são 160 mg/dL e 200 mg/dL;

    14.4%, quando os valores da glicémia são 190 mg/dL e 230 mg/dL.

A partir destes resultados conclui-se que, quando dois indivíduos diferem num determinado valor fixo de glicémia, à medida que aumenta o valor de glicémia dos indivíduos, o que tem o maior valor terá um acréscimo estimado do risco de enfarte cada vez mais pequeno.

4.4.3 Modelo PWP-CP para enfartes múltiplos

O modelo de regressão PWP-CP deve ser usado quando se pretende analisar, para cada acontecimento ocorrido, a influência que uma covariável registada no início da observação do indivíduo tem no tempo

desde esse instante até cada acontecimento. Nesta análise interessa determinar quais os fatores observados na admissão hospitalar que possam influenciar a ocorrência de enfartes múltiplos.

Análise univariável (Modelo PWP-CP)

Na tabela seguinte encontram-se os resultados obtidos pelo ajustamento dos vários modelos PWP-CP aos dados:

Covariáveis	*Valor p*	*RR (IC95%)*
Idade	<0.001	1.038 (1.018;1.059)
IMC	0.884	1.004 (0.951;1.060)
Género	0.404	1.201 (0.781;1.848)
Dislipidémia	0.106	0.715 (0.477;1.074)
Diabetes	0.011	1.703 (1.132;2.563)
Tabagismo	0.391	0.801 (0.484;1.328)
Stress	0.132	0.628 (0.342;1.151)
Ant. Familiares	0.087	0.508 (0.234;1.104)
Hipertensão	0.402	1.207 (0.777;1.873)
AVC/AIT	0.634	1.211 (0.551;2.662)
DAP	0.210	1.719 (0.736;4.015)
Insuf. Cardíaca	0.680	1.096 (0.708;1.697)
TA Sistólica	0.848	0.999 (0.991;1.007)
TA Diastólica	0.325	0.993 (0.979;1.007)
Frequência Cardíaca	0.253	0.993 (0.981;1.005)
Classe KK II	0.144	1.501 (0.871;2.588)
Classe KK III	0.142	2.245 (0.763;6.608)
Classe KK IV	0.532	1.725 (0.313;9.509)
Classe KK I	Classe de referência	
Diag: EAM com supST	0.075	1.941 (0.935;4.030)
Diag: EAM sem supST	0.009	2.555 (1.268;5.152)
Diag: AI	Classe de referência	
Creatinina	0.034	1.193 (1.014;1.404)
TFG	0.004	0.988 (0.980;0.996)
Glicémia	0.005	1.002 (1.001;1.004)
Colesterol Total	0.960	1.000 (0.996;1.004)
Colesterol HDL	0.092	0.983 (0.964;1.003)
Colesterol LDL	0.533	1.002 (0.997;1.007)
Triglicerídeos	0.974	1.000 (0.999;1.001)

As variáveis que, por si só, têm influência significativa no tempo são: a idade, a diabetes, os antece-

dentes familiares, o diagnóstico, a creatinina, a TFG, a glicémia e o colesterol HDL.

Quando se comparam os doentes com AI com os doentes de cada um dos outros diagnósticos em separado, verifica-se um aumento significativo do risco para os doentes diagnosticados com EAM sem supra de ST. O risco estimado de ocorrência de enfarte para os doentes com EAM sem supra ST é 2.555 vezes o risco dos doentes com AI. O risco estimado de ocorrência de enfarte dos doentes com EAM com supra de ST relativamente aos doentes com AI é 1.941. Os doentes diabéticos, por sua vez, apresentam um acréscimo de 70.3% no risco de ocorrência de enfarte em relação aos não diabéticos. O aumento da idade está também associado a um aumento do risco de ocorrência de enfarte, o mesmo se passando com a glicémia. Quanto à TFG, os doentes com valores mais altos tendem a ter um menor risco de enfarte, o que significa que têm um melhor prognóstico.

Análise multivariável (Modelo PWP-CP)

O modelo final PWP-CP para enfartes múltiplos obtido usando o método de seleção de variáveis *stepwise forward* com teste de razão de verosimilhanças é,

Covariáveis	$\hat{\boldsymbol{\beta}}_j$	$EP(\hat{\boldsymbol{\beta}}_j)$	$EP_r(\hat{\boldsymbol{\beta}}_j)$	*Valor p*	*RR (IC95%)*
Idade	0.043	0.010	0.010	<0.001	1.044 (1.023;1.066)
Diabetes	0.587	0.223	0.216	0.006	1.800 (1.179;2.746)
Colesterol HDL	-0.029	0.012	0.010	0.005	0.971 (0.952;0.991)
Colesterol LDL	0.009	0.003	0.003	0.004	1.009 (1.003;1.016)

onde $\mathrm{EP}(\hat{\boldsymbol{\beta}}_j)$ e $\mathrm{EP_r}(\hat{\boldsymbol{\beta}}_j)$ representam a estimativa usual e a estimativa robusta do erro padrão de $\hat{\boldsymbol{\beta}}_j$, respetivamente. Note-se que, quando considerada em conjunto com outras covariáveis, o colesterol LDL passa a ter influência significativa no tempo até enfartes múltiplos. De seguida analisam-se os resíduos do modelo.

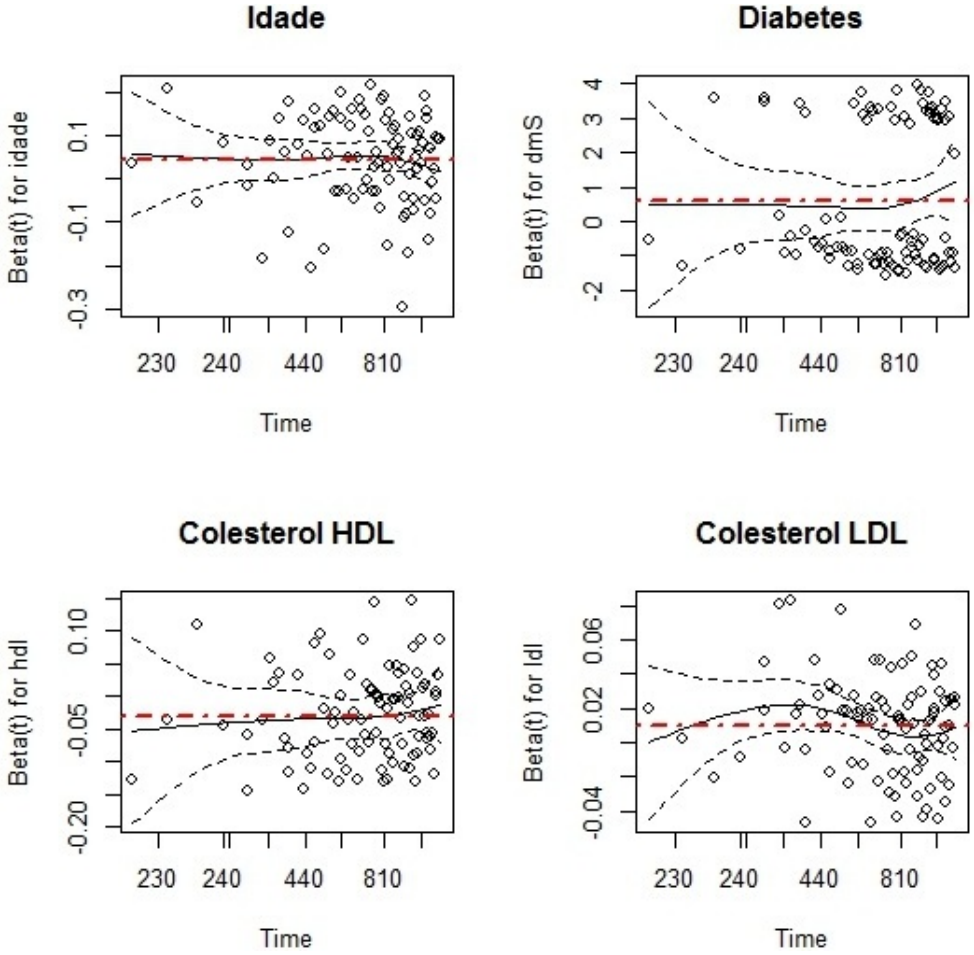
Resíduos de Schoenfeld ponderados (Modelo PWP-CP)

Os resultados do teste da hipótese de proporcionalidade global, assim como para cada uma das covariáveis, podem ser consultados na tabela que se segue:

	rho	*chisq*	*Valor p*
Idade	-0.053	0.279	0.597
Diabetes	0.071	0.456	0.499
Colesterol HDL	0.098	0.434	0.510
Colesterol LDL	-0.142	1.593	0.207
GLOBAL	NA	2.159	0.706

Conclui-se que não existe evidência para rejeitar a hipótese de proporcionalidade global do modelo.

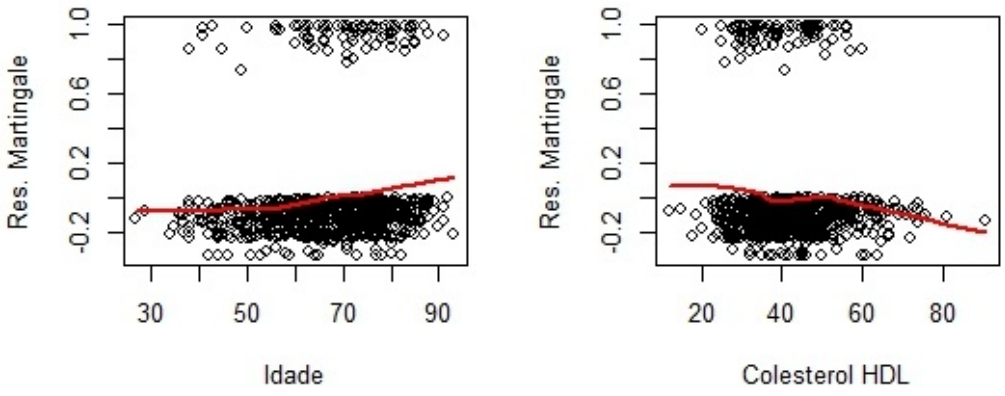
A avaliação da proporcionalidade para as covariáveis é feita também com recurso aos gráficos seguintes:



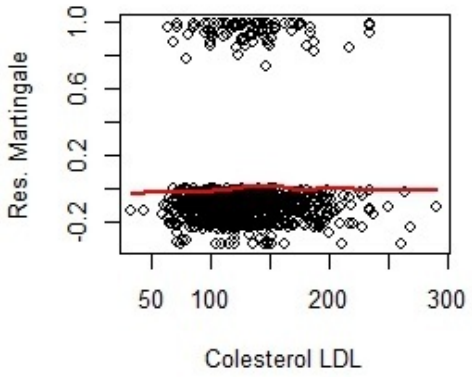
Os efeitos da idade, da diabetes e do HDL são constantes ao longo do tempo. Em relação ao LDL, apesar do comportamento oscilatório da curva de suavização *spline*, o valor estimado do parâmetro mantém-se dentro das curvas do intervalo de confiança, concluindo-se que existe proporcionalidade.

Resíduos martingala (Modelo PWP-CP)

Os gráficos dos resíduos martingala para a avaliação da forma funcional das covariáveis contínuas são:



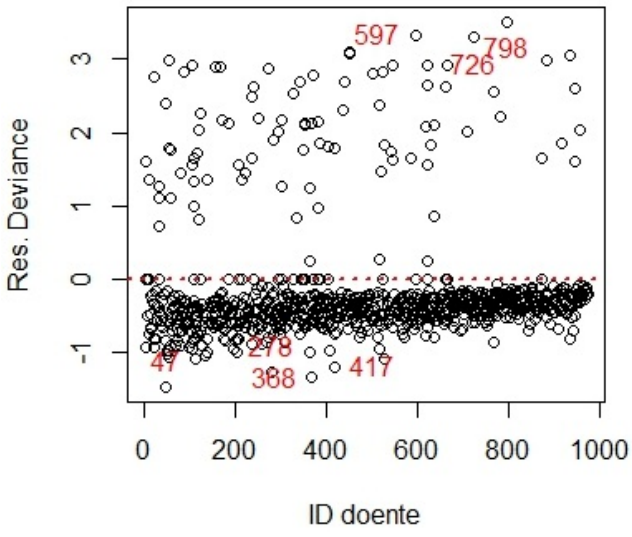
A partir do gráfico da idade não existem dúvidas quanto à forma funcional linear, portanto continua a não ser necessário proceder a uma transformação. O gráfico do HDL é idêntico ao que foi obtido para o modelo de Cox e também aqui se conclui não ser necessário transformar a covariável.



Quanto à covariável LDL, também não existe necessidade de a transformar.

Resíduos *deviance* (Modelo PWP-CP)

O gráfico dos resíduos *deviance* segue abaixo:



Ao observar o gráfico destacam-se os doentes número 47, 278, 368 e 417 com resíduos negativos e os doentes número 597, 726 e 798 com resíduos positivos. Na tabela seguinte apresentam-se, para estes sete doentes, os respetivos valores observados para as covariáveis incluídas no modelo multivariável. Nas colunas "Ordem", "Status" e "Dias" indica-se, respetivamente, qual a ordem da observação considerada *outlier* (primeiro, segundo, terceiro ou quarto tempo observado); quais os tempos até enfarte (S) e quais os censurados (N) e, finalmente, todos os tempos registados para o doente, medidos desde o início do período de observação.

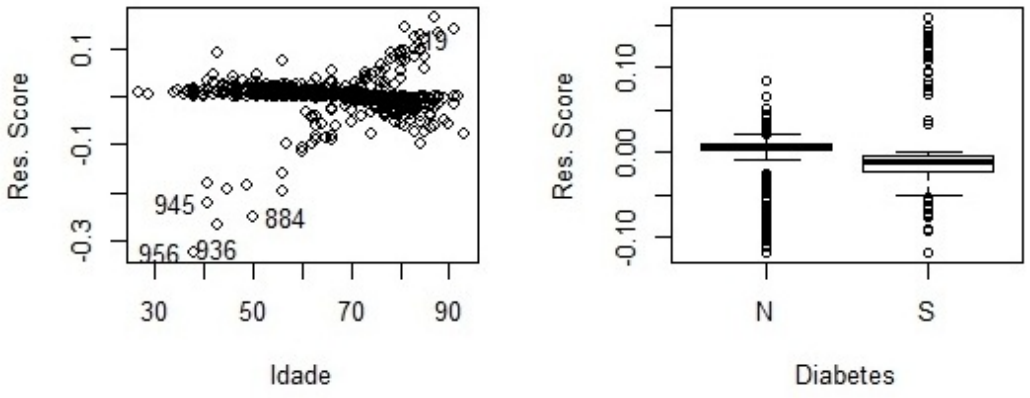
ID doente	*Idade*	*Diab*	*HDL*	*LDL*	*Ordem*	*Status*	*Dias*
47	84	Não	31	88	2.°	S/N	51/677
368	71	Não	28	120	2.°	S/N	10/876
278	74	Sim	27	169	1.°	N/-	846/ -
417	77	Não	30	134	2.°	S/N	160/826
726	60	Sim	44	122	1.°	S/N	5/335
597	65	Sim	49	107	1.°	S/N	9/14
798	56	Não	47	132	1.°	S/N	6/371

Os tempos com resíduos negativos tendem a ser de valor elevado e censurados. Em três dos doentes, o tempo mal ajustado refere-se ao segundo tempo observado. Os quatro doentes tendem a ter uma idade mais avançada, a ter HDL baixo e LDL elevado. Em relação à diabetes, apenas um é diabético. Apesar destas características não sofreram enfarte.

Em relação aos três tempos com resíduo positivo, correspondem todos ao primeiro enfarte ocorrido. Todos os doentes sofreram apenas um enfarte e caracterizam-se por não serem idosos, terem HDL normal e LDL normal. No entanto, sofreram o primeiro enfarte nos primeiros 10 dias após a admissão hospitalar.

Resíduos *score* ponderados (Modelo PWP-CP)

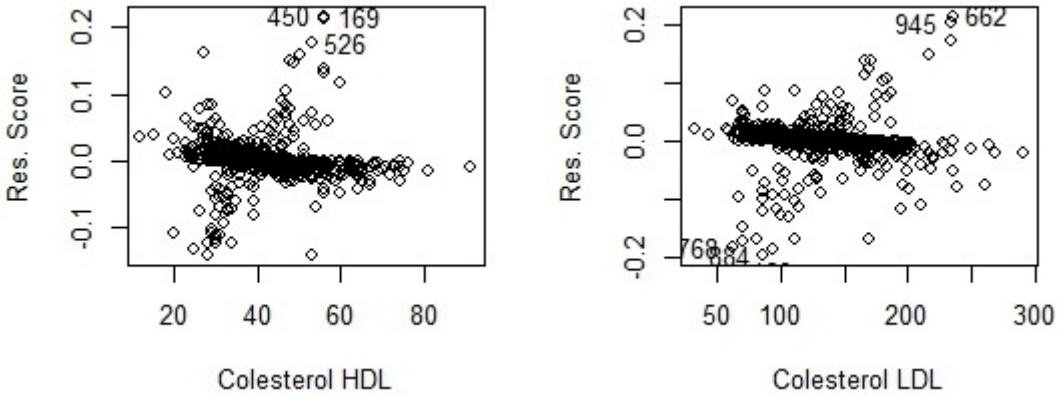
Os pontos influentes podem ser identificados a partir dos gráficos seguintes



Para a diabetes, os valores dos resíduos não são elevados e por isso não se detetam doentes cujos valores foram influentes. Os pontos influentes na estimação do efeito da idade referem-se aos tempos até ao primeiro enfarte observados para os indivíduos 19, 884, 936, 945 e 956.

ID doente	19	884	936	945	956
Idade	87	50	43	41	38
Enfarte	1.°	1.°	1.°	1.°	1.°
Dias	7	66	32	37	491

Neste caso foram mais influentes as observações de doentes que sofreram enfartes poucos dias depois da admissão hospitalar e que tinham idade acima dos 80 anos ou inferior ou igual a 50 anos.
Os resíduos para o colesterol HDL e LDL podem ser analisados a partir dos gráficos seguintes:



Os doentes 169, 450 e 526 são influentes no caso do colesterol HDL.

ID doente	169	450	526
HDL	56	56	53
Enfarte	1.º	1.º	1.º
Dias	258	11	46

Estes têm tempos até enfarte curtos (exceto o doente 169) e valores de HDL próximos de 55, que neste contexto são elevados. Enquanto os pontos influentes, no caso do colesterol LDL, correspondem aos doentes: 169, 662, 768, 884 e 945.

ID doente	169	662	768	884	945
LDL	95	235	63	87	234
Enfarte	1.º	1.º	1.º	1.º	1.º
Dias	258	15	55	66	37

Conclui-se, da tabela anterior, que estes tendem a ter tempos até enfarte curtos e valores de LDL baixos.

Interpretação dos resultados (Modelo PWP-CP)

Após a análise dos resíduos e depois de concluir que o modelo se ajusta satisfatoriamente bem aos dados, pode-se interpretar os resultados obtidos. Pode concluir-se que a idade, a diabetes, o colesterol HDL e o colesterol LDL têm influência significativa no tempo até enfarte.

Os resultados que a seguir são indicados, para cada covariável, referem-se a indivíduos que têm valores iguais das restantes covariáveis.
- Ao compararmos dois indivíduos que diferem apenas num ano de idade, o mais velho tem um aumento

estimado do risco de sofrer um enfarte de 4.4%. Quando se comparam indivíduos cuja diferença de idades é de 5 anos, o mais velho tem um acréscimo de 24.1% no risco de enfarte.
- Estima-se que a presença de diabetes em doentes com SCA aumenta o risco em 80%, quando comparado com os não diabéticos.
- Em relação ao colesterol HDL, considerando dois doentes que difiram em 1 mg/dL, o que apresenta menor valor tem um acréscimo do risco de enfarte estimado em 2.9%. Quando a diferença observada é de 5 mg/dL, o aumento do risco é de 13.5%.
- No que diz respeito ao colesterol LDL na admissão, quando se comparam dois doentes com apenas 1 mg/dL de diferença, o que tem um valor mais alto tem um aumento do risco de sofrer múltiplos enfartes de 0.9%. Quando a diferença é de 5 mg/dL, o acréscimo estimado do risco é de 4.8%.

4.4.4 Modelo PWP-GT para enfartes múltiplos

O modelo de regressão PWP-GT deve ser usado quando se pretende analisar, para cada acontecimento ocorrido, a influência que uma covariável tem no tempo desde o último acontecimento ocorrido, ou desde o início da observação quando ainda não ocorreram acontecimentos. Nesta análise interessa determinar quais os fatores observados na admissão hospitalar que possam influenciar a ocorrência de enfartes múltiplos, desde o último enfarte ou desde o início da observação quando ainda não ocorreram enfartes.
Espera-se que os resultados deste modelo não sejam muito diferentes dos resultados já obtidos a partir do modelo PWP-CP. Apesar de terem definições diferentes para a escala de tempo, os dois modelos PWP têm, para cada acontecimento, a mesma amplitude do intervalo de risco. De facto, as diferenças observadas nas estimativas dos efeitos das covariáveis obtidas a partir de dois modelos, que difiram quanto às amplitudes dos intervalos de risco, são mais marcantes do que as diferenças observadas nos resultados obtidos a partir de dois modelos que difiram quanto às definições da escala de tempo (modelos PWP-CP e PWP-GT).

Análise univariável (Modelo PWP-GT)

Com o intuito de estimar a influência que cada covariável tem, por si só, no tempo até enfarte (desde o último enfarte), ajustaram-se vários modelos PWP-GT univariáveis, cujos resultados são apresentados na tabela seguinte:

Covariáveis	*Valor p*	*RR (IC95%)*
Idade	<0.001	1.036 (1.017;1.056)
IMC	0.990	1.000 (0.951;1.052)
Género	0.503	1.155 (0.757;1.763)
Dislipidémia	0.114	0.729 (0.493;1.079)
Diabetes	0.025	1.566 (1.057;2.319)
Tabagismo	0.490	0.842 (0.517;1.372)
Stress	0.112	0.627 (0.352;1.115)
Ant. Familiares	0.127	0.549 (0.254;1.185)
Hipertensão	0.310	1.246 (0.815;1.905)
AVC/AIT	0.856	1.074 (0.497;2.323)
DAP	0.345	1.515 (0.639;3.592)
Insuf. Cardíaca	0.178	1.302 (0.887;1.911)
TA Sistólica	0.665	0.998 (0.990;1.007)
TA Diastólica	0.304	0.993 (0.979;1.007)
Frequência Cardíaca	0.259	0.993 (0.981;1.005)
Classe KK II	0.074	1.614 (0.956;2.727)
Classe KK III	0.062	2.767 (0.949;8.069)
Classe KK IV	0.444	1.832 (0.389;8.625)
Classe KK I	Classe de referência	
Diag: EAM com supST	0.070	1.918 (0.950;3.876)
Diag: EAM sem supST	0.008	2.488 (1.274;4.858)
Diag: AI	Classe de referência	
Creatinina	0.103	1.145 (0.973;1.347)
TFG	0.004	0.989 (0.981;0.996)
Glicémia	<0.001	1.002 (1.001;1.003)
Colesterol Total	0.811	1.000 (0.997;1.004)
Colesterol HDL	0.024	0.978 (0.959;0.997)
Colesterol LDL	0.572	1.001 (0.996;1.007)
Triglicerídeos	0.932	1.000 (0.999;1.001)

As covariáveis que têm influência significativa no tempo até enfarte são: a idade, a diabetes, a classe KK, o diagnóstico, a TFG, a glicémia e o colesterol HDL.

No que diz respeito ao diagnóstico os resultados são idênticos aos dos modelos de Cox e PWP-CP. Os doentes diabéticos apresentam um acréscimo de 56.6% no risco de ocorrência de enfarte em relação aos não diabéticos. Também neste caso, o aumento da idade ou da glicémia está associado a um aumento do risco de ocorrência de enfarte, enquanto os doentes com valores mais altos de TFG tendem a ter um menor risco de enfarte. Relativamente à classe KK, os doentes com classe IV não diferem quanto ao

risco estimado de enfarte comparativamente aos doentes com classe I. No entanto, os doentes com classe II e III diferem dos doentes com classe I e têm riscos estimados de enfarte 1.614 e 2.767, respetivamente.

Análise multivariável (1º Modelo PWP-GT)

O modelo final PWP-GT obtido usando o método de seleção de variáveis *stepwise forward* com teste de razão de verosimilhanças é

Covariáveis	$\hat{\boldsymbol{\beta}}_j$	$EP(\hat{\boldsymbol{\beta}}_j)$	$EP_r(\hat{\boldsymbol{\beta}}_j)$	*Valor p*	*RR (IC95%)*
Idade	0.041	0.010	0.010	<0.001	1.042 (1.021;1.062)
Colesterol HDL	-0.035	0.012	0.010	0.001	0.966 (0.946;0.986)
Colesterol LDL	-0.009	0.003	0.003	0.006	1.009 (1.003;1.015)
Glicémia	0.002	0.001	0.001	0.001	1.002 (1.001;1.004)

onde $\mathrm{EP}(\hat{\boldsymbol{\beta}}_j)$ e $\mathrm{EP_R}(\hat{\boldsymbol{\beta}}_j)$ representam a estimativa usual e a estimativa robusta do erro padrão de $\hat{\boldsymbol{\beta}}_j$, respetivamente.

A partir da análise dos resíduos de Schoenfeld conclui-se que os riscos associados às quatro covariáveis são proporcionais. Quanto à forma funcional das quatro covariáveis contínuas, à semelhança do que aconteceu no caso do modelo de Cox, também aqui é necessário proceder a uma transformação da glicémia. Os gráficos dos resíduos de Schoenfeld e dos resíduos martingala podem ser consultados nos apêndices D.2 e D.3, respetivamente. De entre as várias transformações possíveis, também nesta análise se optou pelo inverso da glicémia.

Análise multivariável (2.º Modelo PWP-GT)

O modelo PWP-GT multivariável obtido a partir do método de seleção de variáveis *stepwise forward* que considerou a transformada da glicémia em vez da covariável original é

Covariáveis	$\hat{\boldsymbol{\beta}}_j$	$EP(\hat{\boldsymbol{\beta}}_j)$	$EP_r(\hat{\boldsymbol{\beta}}_j)$	*Valor p*	*RR (IC95%)*
Idade	0.040	0.010	0.010	<0.001	1.041 (1.021;1.062)
Colesterol HDL	-0.034	0.012	0.010	0.001	0.966 (0.947;0.986)
Colesterol LDL	-0.009	0.003	0.003	0.007	1.009 (1.002;1.015)
1/Glicémia	-100.20	45.28	45.72	0.028	(*)

(*) *RR(IC95%)*: 3.186×10^{-44} (3.882×10^{-83};2.614×10^{-5}).

Note-se que, quando considerada em conjunto com outras covariáveis, o colesterol LDL passa a ter influência significativa no tempo até enfartes múltiplos.

As covariáveis incluídas neste modelo são as mesmas do primeiro modelo PWP-GT, tendo-se substituído apenas a glicémia pela sua transformada. No entanto, terá de se repetir a análise dos resíduos

de Schoenfeld e martingala para todas as covariáveis.

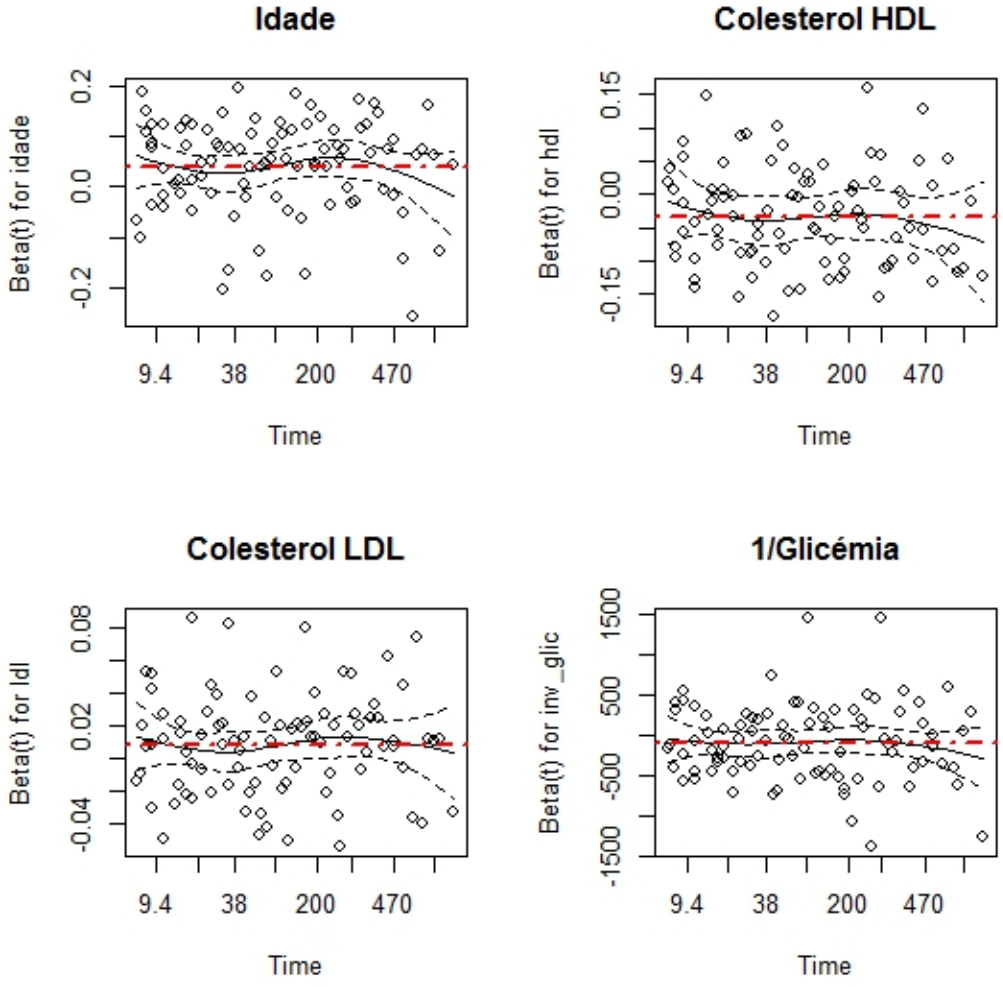
Resíduos de Schoenfeld ponderados (2.º Modelo PWP-GT)

Os resultados do teste da hipótese de proporcionalidade são:

	rho	*chisq*	*Valor p*
Idade	-0.044	0.174	0.676
Colesterol HDL	-0.101	0.529	0.467
Colesterol LDL	0.028	0.069	0.793
1/Glicémia	-0.046	0.231	0.631
GLOBAL	NA	0.964	0.915

Na globalidade o modelo não viola a hipótese de proporcionalidade das funções de risco.

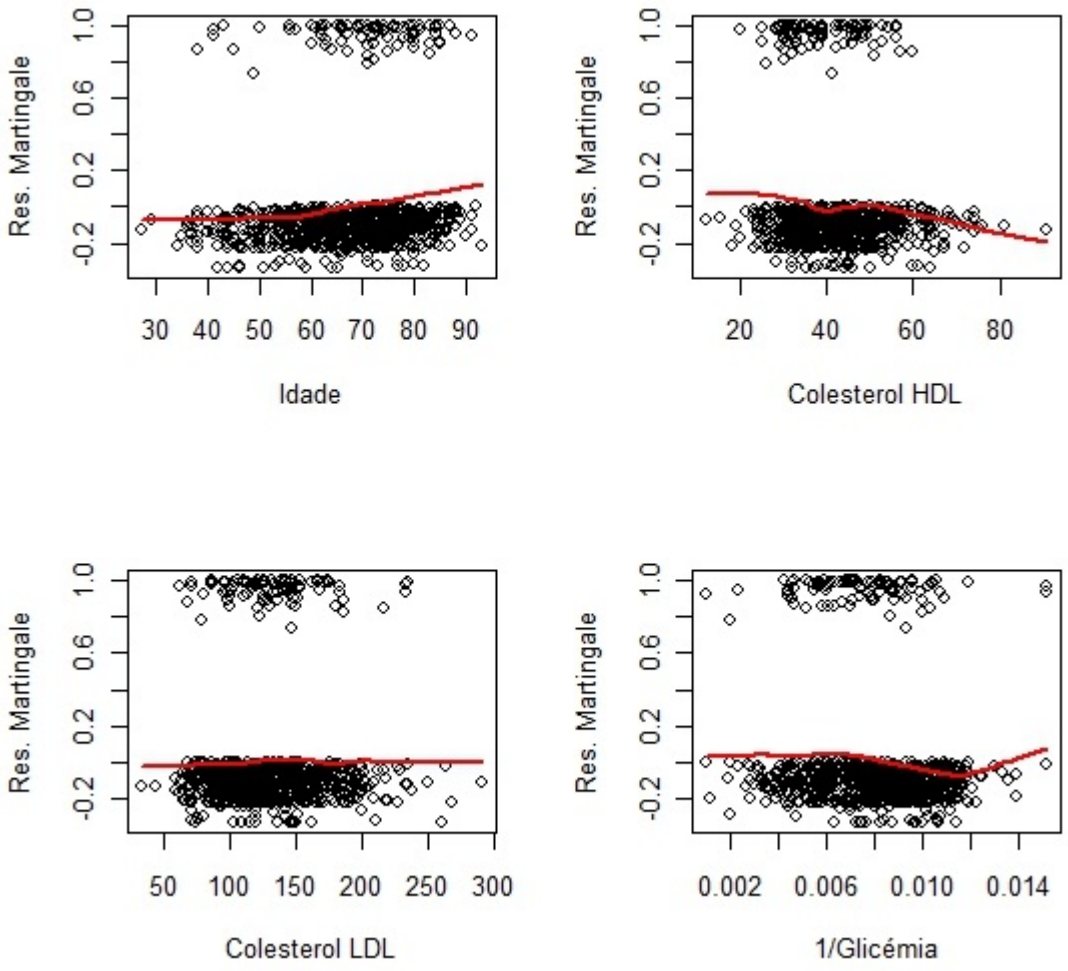
A partir dos gráficos dos resíduos de Schoenfeld ponderados para cada uma das covariáveis,



conclui-se que a idade, o colesterol HDL, o colesterol LDL e a transformada da glicémia têm um efeito constante no tempo até enfarte.

Resíduos martingala (2.º Modelo PWP-GT)

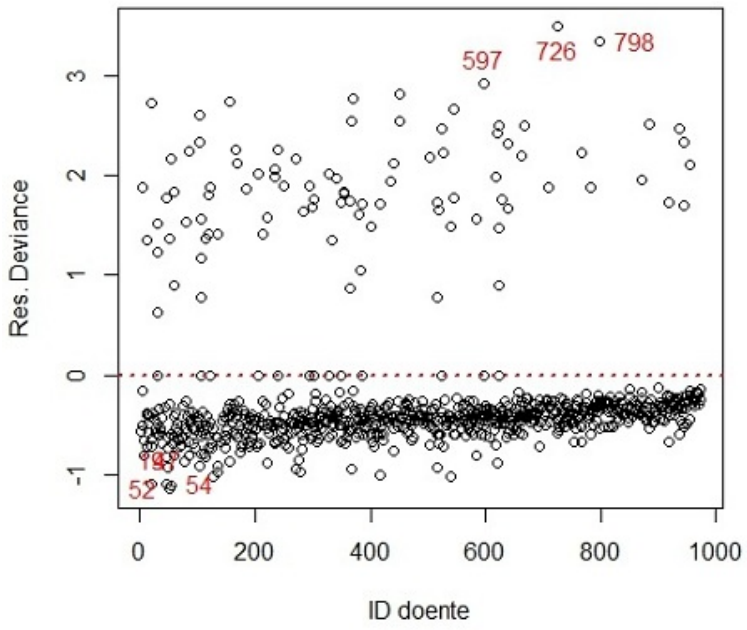
Os gráficos dos resíduos martingala para cada covariável contínua são os seguintes:



Os gráficos das covariáveis idade, HDL e LDL permitem confirmar que não existe necessidade de proceder a uma transformação. Em relação à transformada da glicémia, observam-se algumas oscilações, sendo a mais marcada a partir do valor 0.011. No entanto, não deve ser valorizada por existirem poucas observações acima desse valor e existirem dois valores superiores a 0.014 com resíduo martingala próximo de 1. Assim, confirma-se que a transformação da glicémia foi adequada.

Resíduos *deviance* (2.º Modelo PWP-GT)

O gráfico dos resíduos *deviance* segue abaixo:



Ao observar o gráfico destacam-se os doentes número 19, 47, 52 e 54 com resíduos negativos e os doentes número 597, 726 e 798 com resíduos positivos. Na tabela seguinte apresentam-se, para estes sete doentes, os respetivos valores observados para as covariáveis incluídas no modelo multivariável e todos os tempos registados desde o último enfarte, ou desde o início do período de observação quando se trata do primeiro tempo observado.

ID doente	*Idade*	*HDL*	*LDL*	*Glic*	*Ordem*	*Status*	*Dias*
19	87	47	142	169	2.°	S/N	7/436
47	84	31	88	118	2.°	S/N	51/626
52	84	30	117	135	2.°	S/N	187/434
54	84	34	86	141	2.°	S/N	14/548
597	65	49	107	147	1.°	S/N	9/5
726	60	44	122	140	1.°	S/N	5/330
798	56	47	132	139	1.°	S/N	6/365

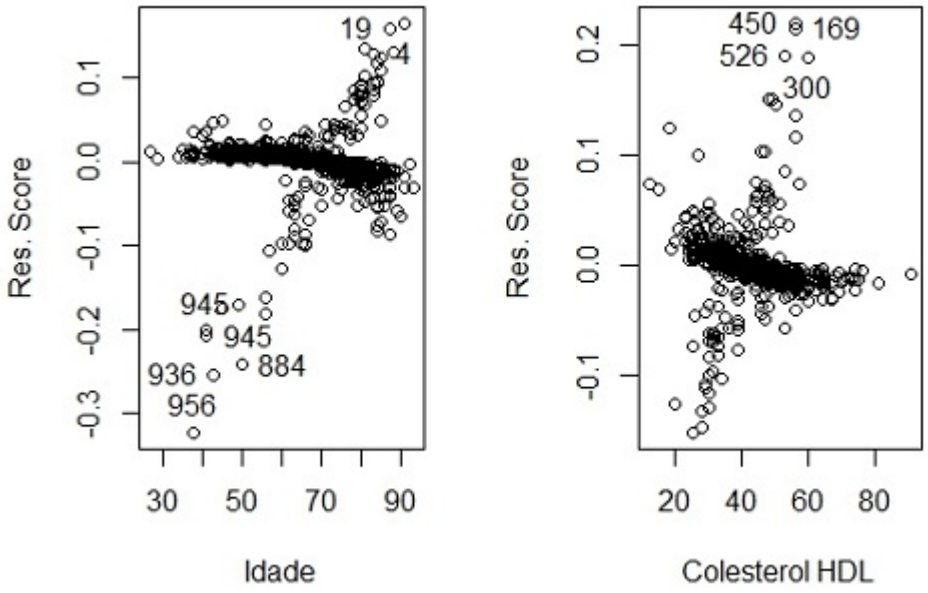
Os tempos com resíduos negativos referem-se todos a tempos censurados, de valor elevado e medidos desde o primeiro enfarte. Os quatro doentes tendem a ter uma idade mais avançada, a ter HDL baixo e LDL elevado. Todos os doentes têm metabolismo da glicémia anormal (superior a 110 mg/dL) e três deles têm glicémia superior a 126 mg/dL (limite a partir do qual se diagnostica a diabetes). Apesar destas características não sofreram enfarte.

Em relação aos três tempos com resíduo positivo, correspondem todos ao primeiro enfarte ocorrido. Os doentes, para os quais se observaram estes tempos, caracterizam-se por não serem idosos, terem HDL normal e LDL inferior a 135 mg/dL. Apesar de terem valores de glicémia acima de 126 mg/dL

considera-se que sofreram o primeiro enfarte precocemente.

Resíduos *score* ponderados (2.º Modelo PWP-GT)

Os pontos influentes para a idade e o HDL podem ser identificados a partir dos gráficos seguintes:



Os pontos influentes na estimação do efeito da idade referem-se aos tempos observados para os indivíduos 4, 19, 884, 936, 945 (dois tempos) e 956.

ID doente	4	19	884	936	945	956
Idade	91	87	50	43	41	38
Enfarte	1.º	1.º	1.º	1.º	1.º/2.º	1.º
Dias	39	7	66	32	37/167	491

Estes doentes sofreram enfarte poucos dias depois da admissão hospitalar (ou desde o primeiro enfarte, no caso do indivíduo 945) e têm idade acima dos 80 anos ou inferior ou igual a 50 anos. Note-se que o primeiro tempo observado para o doente 19 é influente, no entanto, a partir da análise dos resíduos *deviance*, concluiu-se que o seu segundo tempo é um *outlier*.

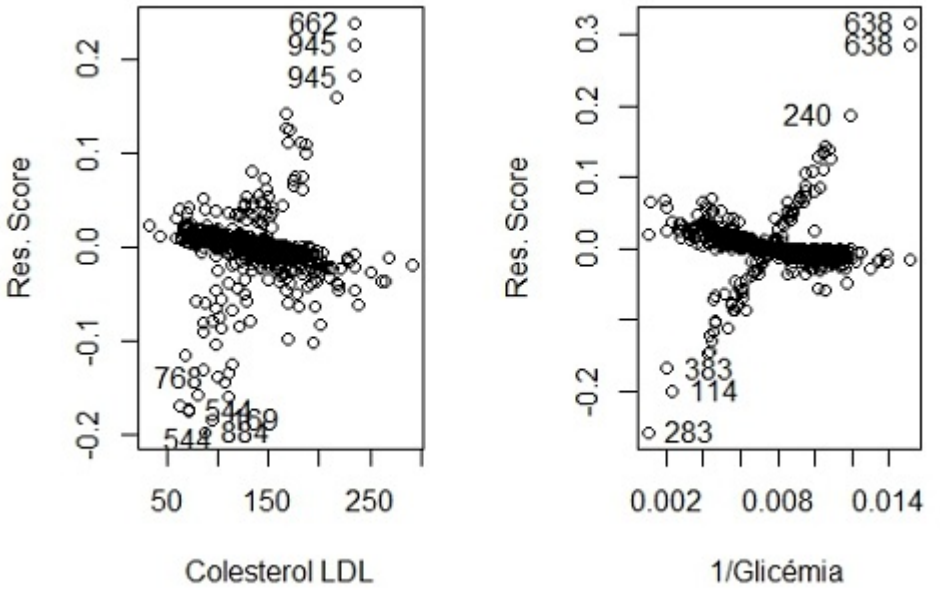
Os pontos influentes na estimação do efeito do HDL referem-se aos tempos até ao primeiro enfarte observados para os indivíduos 169, 300, 450 e 526.

ID doente	169	300	450	526
HDL	56	60	56	53
Enfarte	1.º	1.º	1.º	1.º
Dias	258	469	11	46

Estes doentes têm valores de HDL próximos de 55, que neste contexto são elevados. Os tempos até enfarte para os doentes 169 e 300 não são curtos, o que leva a concluir que estas duas observações são

influentes pelo seu valor de HDL.

Os resíduos para o colesterol LDL e a transformada da glicémia podem ser analisados a partir dos gráficos seguintes:



Os tempos influentes, no caso do colesterol LDL, correspondem aos doentes: 169, 544 (dois tempos), 662, 768, 884 e 945 (dois tempos).

ID doente	169	544	662	768	884	945
LDL	95	71	235	63	87	234
Enfarte	1.°	1.°/2.°	1.°	1.°	1.°	1.°/2.°
Dias	258	10/137	15	55	66	37/167

Conclui-se, da tabela anterior, que estas observações estão associadas a valores de LDL abaixo de 100 mg/dL ou acima de 230 mg/dL.

Os tempos influentes na estimativa do efeito da transformada da glicémia foram observados para os doentes: 114, 240, 283, 383 e 638 (dois tempos).

ID doente	114	240	283	383	638
Glicémia	427	84	998	495	66
Enfarte	1.°	1.°	1.°	1.°	1.°/2.°
Dias	207	40	278	813	82/300

Os doentes 240 e 638 são influentes porque têm valores de glicémia mais baixos e sofreram enfarte precocemente (exceto o segundo enfarte do doente 638). Quanto aos restantes doentes a sua influência deve-se ao facto do valor da glicémia ser bastante elevado (muito acima do valor 126 mg/dL a partir do qual se diagnostica a diabetes).

Interpretação dos resultados (2.º Modelo PWP-GT)

Após a análise dos resíduos e depois de concluir que o modelo ajustado é adequado, está-se em condições de interpretar os resultados obtidos. Pode concluir-se que a idade, o colesterol HDL, o colesterol LDL e a glicémia têm influência significativa no tempo.

Os resultados que a seguir são indicados, para cada covariável, referem-se a indivíduos que têm valores iguais das restantes covariáveis.

- Ao compararmos dois indivíduos que diferem apenas num ano de idade, o mais velho tem um aumento estimado do risco de sofrer um enfarte de 4.1%. Quando se comparam indivíduos cuja diferença de idades é de 5 anos, o mais velho tem um acréscimo de 22.3% no risco de enfarte.
- Em relação ao colesterol HDL, considerando dois doentes que difiram em 1 mg/dL, o que apresenta menor valor tem um acréscimo do risco de enfarte estimado em 3.4%. Quando a diferença observada é de 5 mg/dL, o aumento do risco é de 15.7%.
- No que diz respeito ao colesterol LDL na admissão, quando se comparam dois doentes com apenas 1 mg/dL de diferença, o que tem um valor mais alto tem um aumento do risco de sofrer múltiplos enfartes de 0.9%. Quando a diferença é de 5 mg/dL, o acréscimo estimado do risco é de 4.5%.
- A interpretação da influência da glicémia no tempo até enfarte depende dos valores considerados para os dois doentes que se pretende comparar. Quando diferem, em 40 mg/dL, o aumento estimado do risco é

    33.1%, quando os valores da glicémia são 100 mg/dL e 140 mg/dL;

    19.9%, quando os valores da glicémia são 130 mg/dL e 170 mg/dL;

    13.3%, quando os valores da glicémia são 160 mg/dL e 200 mg/dL;

    9.6%, quando os valores da glicémia são 190 mg/dL e 230 mg/dL.

Conclui-se que, quando dois indivíduos diferem num determinado valor fixo de glicémia, à medida que aumenta o valor de glicémia dos indivíduos, o que tem o maior valor terá um acréscimo estimado do risco de enfarte cada vez mais pequeno.

4.5 Conclusões e trabalho futuro

A partir dos resultados obtidos pelo ajustamento dos modelos de Cox, PWP-CP e PWP-GT pode concluir-se que:

- A frequência cardíaca e a insuficiência cardíaca apenas têm influência significativa na ocorrência do primeiro enfarte. Como não foram incluídas nos modelos para enfartes múltiplos, conclui-se que não

têm um efeito significativo quando se consideram enfartes recorrentes;
- O colesterol LDL, por seu lado, não tem efeito na ocorrência apenas do primeiro enfarte, no entanto, tem influência na ocorrência de vários enfartes. Quando se comparam indivíduos cuja diferença de colesterol LDL é de 5 mg/dL, o que apresenta um valor mais alto tem um acréscimo de 4.8% no risco de enfarte quando se considera o tempo desde a admissão hospitalar e tem um acréscimo de 4.5% no risco de enfarte quando se considera o tempo desde o último enfarte;
- A idade e o colesterol HDL são covariáveis que influenciam significativamente o tempo até à ocorrência do primeiro enfarte, assim como o tempo até à ocorrência de enfartes múltiplos.
Quando se comparam indivíduos cuja diferença de idades é de 5 anos, conclui-se que o mais velho tem um acréscimo de 18.9% no risco do primeiro enfarte; tem um acréscimo de 24.1% no risco de enfarte quando se considera o tempo desde a admissão hospitalar e tem um acréscimo de 22.3% no risco de enfarte quando se considera o tempo desde o último enfarte;
Quanto ao colesterol HDL, quando se comparam indivíduos cuja diferença é de 5 mg/dL, o que apresenta um valor mais baixo tem um acréscimo de 11.2% no risco de enfarte quando se considera o tempo até ao primeiro enfarte; tem um acréscimo de 13.5% no risco de enfartes múltiplos quando se considera o tempo desde a admissão hospitalar e tem um acréscimo de 15.7% no risco de enfartes múltiplos quando se considera o tempo desde o último enfarte;
- O valor da glicémia tem influência significativa no tempo até ao primeiro enfarte e no tempo até enfartes múltiplos quando medido desde o último enfarte, enquanto a diabetes tem influência significativa no tempo até enfartes múltiplos desde o início da observação. Estas duas covariáveis estão de certa forma relacionadas porque se referem ao metabolismo da glicose. A diabetes diagnostica-se a partir do valor de glicémia. Um diabético controlado tende a apresentar valor de glicémia razoavelmente normal, enquanto um diabético não controlado pode apresentar valores anormais. Assim, a diabetes tem influência significativa no tempo até enfartes múltiplos, medido desde a admissão hospitalar, independentemente do valor de glicémia registado. No caso do tempo até ao primeiro enfarte e do tempo desde o último enfarte, é o valor de glicémia registado na admissão hospitalar que tem influência no tempo e não a presença de diabetes;

Conclui-se que, como era de esperar, os resultados dos modelos PWP-CP e PWP-GT são idênticos, diferindo apenas na inclusão da diabetes no modelo PWP-CP em vez da glicémia que foi incluída no modelo PWP-GT.

É ainda importante referir que a partir do ajustamento dos modelos PWP-CP e PWP-GT se obtiveram estimativas robustas do erro padrão dos coeficientes inferiores às estimativas usuais, o que leva a

concluir que existe maior variabilidade entre os tempos observados para o mesmo indivíduo, do que entre os tempos observados para indivíduos diferentes.

Trabalho futuro

Quando se realizam estudos em que estão envolvidos acontecimentos recorrentes, sabe-se que a correlação observada entre os tempos se pode dever a:

Heterogeneidade entre indivíduos: alguns indivíduos têm taxa global de recorrência diferente dos outros devido a fatores desconhecidos, não medidos ou não mensuráveis, como por exemplo, o estilo de vida ou o código genético. Como consequência, os acontecimentos tendem a ocorrer mais precocemente a uns indivíduos e mais tardiamente a outros. Estes fatores introduzem heterogeneidade entre os indivíduos e produzem correlação entre os acontecimentos observados para o mesmo indivíduo, que se manifesta tanto pelas diferentes taxas de ocorrência global como pelas diferenças observadas nos tempos para o mesmo indivíduo;

Dependência entre acontecimentos: a ocorrência de um acontecimento pode tornar a ocorrência de acontecimentos futuros mais ou menos provável. Esta dependência entre acontecimentos pode ser produzida por um enfraquecimento biológico ("deterioração") ou por um fortalecimento biológico ("resistência"). Qualquer um destes fenómenos implica que o risco associado a um acontecimento está dependente dos acontecimentos ocorridos no passado, o que também provoca correlação entre acontecimentos observados para o mesmo indivíduo.

A investigação médica e a prática clínica sugerem que tanto a heterogeneidade como a dependência entre os acontecimentos encontradas nos estudos de acontecimentos recorrentes serão a regra e não a excepção. Várias extensões do modelo de Cox têm sido amplamente utilizadas na modelação de acontecimentos recorrentes, como descrito no capítulo 3. Seria interessante, como trabalho futuro, comparar vários modelos, no contexto dos acontecimentos recorrentes, que levem em conta a dependência entre acontecimentos e a heterogeneidade entre indivíduos. Um dos modelos que poderá estudar-se será o modelo com fragilidade condicional com tempo definido por intervalos (*gap time*), que incorpora um efeito aleatório para modelar a heterogeneidade e considera uma variável de estratificação e um conjunto de indivíduos em risco restritivo, para definir a estrutura de dependência condicional dos acontecimentos (Box-Steffensmeier e De Boef, 2006).

Apêndice A

Método de seleção de variáveis

Nesta secção vai abordar-se a questão da seleção das variáveis a incluir no modelo. Se se tiver observado um grande número de variáveis, não é geralmente possível incluí-las todas no modelo e um dos motivos é o facto de se pretender que o modelo de regressão seja parcimonioso. Dos vários métodos de seleção existentes apenas será ilustrado o método *stepwise*.

Este método pode usar qualquer um dos testes da secção 2.5.1 para testar a inclusão de covariáveis no modelo. A escolha do teste não é importante visto os resultados dos diferentes testes, na maioria das situações, serem muito semelhantes. Nas situações em que não são semelhantes, deve optar pelo teste de razão de verosimilhanças. Por esse motivo será este o teste usado na descrição do método de seleção de variáveis *stepwise*.

Existem três algoritmos *stepwise* de seleção de variáveis. O mais comum consiste na inclusão de variáveis pelo método *forward* seguido de uma eliminação de variáveis pelo método *backward*. Os outros dois consistem ou só na aplicação do método *forward* ou só do método *backward*.

Antes de iniciar o algoritmo, é necessário fixar os níveis de significância de entrada e saída de covariáveis no modelo. Define-se p_E como nível de entrada e p_S como nível de saída, onde $p_E \leq p_S$.

Dada uma amostra de dimensão n com q covariáveis observadas, os passos do algoritmo são:

Passo 0:

Calcula-se o valor da log-verosimilhança parcial para o modelo nulo $l(\mathbf{0})$, que não inclui covariáveis. De seguida determinam-se as q log-verosimilhanças parciais correspondentes aos modelos de Cox incluindo apenas cada uma das covariáveis, $l^{(0)}(\mathbf{z}_j), j = 1, \cdots, q$. A estatística de teste para a comparação entre cada um dos modelos e o modelo nulo é,

$$G^{(0)}(\mathbf{z}_j) = -2[l^{(0)}(\mathbf{z}_j) - l(\mathbf{0})], \qquad j = 1, \cdots, q,$$

onde $^{(0)}$ indica que se está no passo zero e j refere que a verosimilhança parcial foi calculada para o modelo de Cox apenas com a covariável z_j.

Para cada covariável z_j, $(j = 1, \cdots, q)$ calcula-se o valor $p^{(0)}(\mathbf{z}_j) = \mathrm{P}[\chi^2_v \geq G^{(0)}(\mathbf{z}_j)]$.

Escolhe-se a variável z_{e1} tal que: $p^{(0)}(\mathbf{z}_{e1}) = \min_j\{p^{(0)}(\mathbf{z}_j)\}$ e $p^{(0)}(\mathbf{z}_{e1}) < p_E$. Se $\forall j \in \{1, \cdots, q\},\ p^{(0)}(\mathbf{z}_j) \geq p_E$ então o algoritmo pára, senão continua para o passo 1.

Passo 1:

Este passo começa com a inclusão no modelo da variável escolhida no passo anterior, z_{e1}, e com a obtenção da sua log-verosimilhança maximizada. Posteriormente constroem-se os $q-1$ modelos de Cox em que se incluem duas covariáveis: a variável z_{e1} e cada uma das variáveis z_j que ainda estão fora do modelo, tal que $j = 1, \cdots, q$ e $j \neq e1$. Determinam-se as log-verosimilhanças correspondentes e obtêm-se as $q-1$ estatísticas de teste e os respetivos valores p,

$$G^{(1)}(\mathbf{z}_j) = -2[l^{(1)}(\mathbf{z}_j) - l(\mathbf{z}_{e1})], \qquad j = 1, \cdots, q \text{ e } j \neq e1$$

$$p^{(1)}(\mathbf{z}_j) = \mathrm{P}[\chi^2_v \geq G^{(1)}(\mathbf{z}_j)], \qquad j = 1, \cdots, q \text{ e } j \neq e1$$

A covariável $\mathbf{z}_{e2}$ que satisfaz as condições $p^{(1)}(\mathbf{z}_{e2}) = \min_{j \neq e1}\{p^{(1)}(\mathbf{z}_j)\}$ e $p^{(1)}(\mathbf{z}_{e2}) < p_E$, será a próxima a incluir no modelo. Se $\forall j \in \{1, \cdots, p\} \setminus \{e1\},\ p^{(1)}(\mathbf{z}_j) \geq p_E$ então o algoritmo pára, senão continua para o passo 2.

Passo 2:

Com a inclusão da variável z_{e2} no modelo, interessa saber se a covariável z_{e1} continua a ter um efeito significativo no tempo de vida. O passo 2 inicia-se com a avaliação da eliminação *backward* de z_{e1}. Assim, começa-se por obter as log-verosimilhanças maximizadas do novo modelo e dos modelos em que são removidas cada uma das variáveis exceto a que entrou no final do passo anterior.

$$G^{(2)}(\mathbf{z}_{e1}) = -2[l^{(2)}(\mathbf{z}_{e1}, \mathbf{z}_{e2}) - l^{(2)}(\mathbf{z}_{e1})]$$

$$p^{(2)}(\mathbf{z}_{e1}) = \mathrm{P}[\chi^2_v \geq G^{(2)}(\mathbf{z}_{e1})]$$

Se $p^{(2)}(\mathbf{z}_{e1}) \geq p_S$ então os modelos com e sem z_{e1} são idênticos e por isso escolhe-se o mais parcimonioso, ou seja, sem z_{e1}. Assumindo que z_{e1} se mantém no modelo, testa-se a entrada de mais covariáveis.

$$G^{(2)}(\mathbf{z}_j) = -2[l^{(2)}(\mathbf{z}_j) - l(\mathbf{z}_{e1}, z_{e2})], \qquad j = 1, \cdots, q \text{ e } j \neq e1, e2$$

$$p^{(2)}(\mathbf{z}_j) = \mathrm{P}[\chi^2_v \geq G^{(2)}(\mathbf{z}_j)], \qquad j = 1, \cdots, q \text{ e } j \neq e1, e2$$

A próxima covariável $\mathbf{z}_{e3}$ a incluir no modelo tem de satisfazer $p^{(2)}(\mathbf{z}_{e3}) = \min_{j \neq e1,e2}\{p^{(2)}(\mathbf{z}_j)\}$ e $p^{(2)}(\mathbf{z}_{e3}) < p_E$. Se $\forall j \in \{1, \cdots, p\} \setminus \{e1, e2\},\ p^{(2)}(\mathbf{z}_j) \geq p_E$ então o algoritmo pára, senão continua para o passo 3.

Passo 3:

Neste passo aplica-se exatamente o mesmo procedimento que foi aplicado no passo 2. No caso de sair alguma covariável tem de se ajustar o modelo sem essa covariável e testa-se uma nova saída. Se não sair mais nenhuma, então está-se em condições de testar a entrada de uma nova covariável.

O processo termina quando num passo m todas as variáveis fora do modelo apresentam $p^{(m)}(\mathbf{z}_j) \geq p_E$ e nenhuma das incluídas no modelo poder ser removida.

Os valores de p_E e p_S podem, por exemplo, ser iguais a 0.05 e 0.10 respetivamente. O número de graus de liberdade do χ^2, representado por v na determinação do valor $p^{(m)}$, é igual à diferença entre o número de parâmetros dos dois modelos envolvidos na comparação. Quando se testa a remoção ou inclusão de uma covariável categórica com k categorias, $v = k - 1$.
Após a conclusão deste algoritmo avalia-se a inclusão de interações entre as covariáveis que foram incluídas no modelo. As interações a estudar devem fazer sentido do ponto de vista clínico e só devem ser incluídas no modelo caso venham a melhorá-lo significativamente.

Apêndice B

Processos de contagem

Com a introdução por Aalen, em meados dos anos 70 do século XX, da teoria dos processos de contagem e das martingalas na análise de sobrevivência, foi possível solidificar a base teórica de muitos dos estimadores existentes na altura, nomeadamente do estimador para o modelo de riscos proporcionais de Cox que tinha sido apresentado em 1972.
Na formulação dos processos de contagem o par (T_i^*, δ_i), onde $T_i^* = \min(T_i, C_i)$ e $\delta_i = \mathrm{I}(\{T_i \leq C_i\})$, é substituído pelo par de funções $(\mathrm{N}_i(t), \mathrm{Y}_i(t))$, definidas por:

- $\mathrm{N}_i(t) = \mathrm{N}^{\mathrm{o}}$ de acontecimentos ocorridos ao indivíduo i no intervalo de tempo $[0, t]$;
- $\mathrm{Y}_i(t) =$ Indica se o indivíduo i está em observação e está no grupo de risco no instante t;

Esta definição permite generalizar a notação usada para situações em que se observam acontecimentos múltiplos. No caso particular em que apenas se observa um acontecimento e existe censura à direita a notação pode ser simplificada:

- $\mathrm{N}_i(t) = \mathrm{I}(\{T_i \leq t, \delta_i = 1\})$ - que toma apenas os valores 0 e 1. Toma o valor 1 quando o acontecimento de interesse já ocorreu e o instante t é superior ou igual ao seu tempo de vida;
- $\mathrm{Y}_i(t) = \mathrm{I}(\{T_i \geq t\})$ - o indivíduo está em risco imediatamente antes do instante t, quando ainda não ocorreu o acontecimento $(T_i \geq t)$ e ainda não ocorreu a censura $(C_i \geq t)$. Quando $t > T_i$ ou $t > C_i$ o indivíduo deixa de estar em risco por já ter saído do estudo.

Exemplos de processos de contagem para acontecimentos múltiplos:

1. O indivíduo entra no estudo no instante 0 e sofre acontecimentos múltiplos nos instantes 1,5 e 2. Sai do estudo no instante 3:

$$\mathrm{N}(t) = \begin{cases} 0 & \text{se} \quad t < 1,5 \\ 1 & \text{se} \quad 1,5 \leq t < 2 \\ 2 & \text{se} \quad t \geq 2. \end{cases} \qquad \text{e} \qquad \mathrm{Y}(t) = \begin{cases} 1 & \text{se} \quad t \leq 3 \\ 0 & \text{se} \quad t > 3. \end{cases}$$

2. O indivíduo não está em risco até ao instante 2 e sofre acontecimentos múltiplos nos instantes 3 e 3,5. Sai do estudo no instante 5:

$$\mathrm{N}(t) = \begin{cases} 0 & \text{se} \quad t < 3 \\ 1 & \text{se} \quad 3 \leq t < 3,5 \\ 2 & \text{se} \quad t \geq 3,5. \end{cases} \qquad \text{e} \qquad \mathrm{Y}(t) = \begin{cases} 0 & \text{se} \quad t \leq 2 \\ 1 & \text{se} \quad 2 < t \leq 5 \\ 0 & \text{se} \quad t > 5. \end{cases}$$

É de notar que o processo $\mathrm{Y}(t)$ é contínuo à esquerda e é previsível, por se saber o seu valor imediatamente antes do instante t, isto é, em t^-. É necessário o indivíduo estar em risco imediatamente antes do instante t para que o acontecimento possa ser observado a ocorrer em t. Por seu lado, o processo $\mathrm{N}(t)$ é contínuo à direita, porque o acontecimento só pode ocorrer exatamente no instante t.
A partir dos processos $\mathrm{N}_i(t)$ e $\mathrm{Y}_i(t)$ define-se:

- $\overline{\mathrm{Y}}(t) = \sum_i \mathrm{Y}_i(t)$ - o número de indivíduos em risco no instante t, mais precisamente no intervalo de tempo infinitesimal $(t-\epsilon, t]$;
- $\overline{\mathrm{N}}(t) = \sum_i \mathrm{N}_i(t)$ - o número total de acontecimentos ocorridos até ao instante t;
- $\triangle\overline{\mathrm{N}}(t) = \overline{\mathrm{N}}(t) - \overline{\mathrm{N}}(t^-)$ - o número de acontecimentos ocorridos exatamente no instante t.

Os estimadores definidos na secção 1.4 podem ser escritos usando a notação dos processos de contagem,

1. $\widehat{S}_{KM}(t) = \prod\limits_{j:x_j \leq t} \left[1 - \triangle\overline{\mathrm{N}}(x_j)/\overline{\mathrm{Y}}(x_j)\right]$;
2. $\widehat{\Lambda}_{NA}(t) = \sum\limits_{j:x_j \leq t} \left[\triangle\overline{\mathrm{N}}(x_j)/\overline{\mathrm{Y}}(x_j)\right]$;
3. $\widehat{S}_B(t) = \prod\limits_{j:x_j \leq t} \mathrm{e}^{-\left[\triangle\overline{\mathrm{N}}(x_j)/\overline{\mathrm{Y}}(x_j)\right]}$.

Dada uma amostra de dimensão n, o modelo de Cox referente à função de risco para o tempo de vida associado a um vetor de p covariáveis possivelmente dependentes do tempo, $\mathbf{z}(t) = (\mathbf{z}_1(t), \cdots, \mathbf{z}_p(t))'$ é dado por, $\lambda(t;\mathbf{z}) = \lambda_0(t)\mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}(t)}$. A função de verosimilhança parcial segundo os processos de contagem é,

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \prod_{t \geq 0} \left\{ \frac{\mathrm{Y}_i(t)\mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_i(t)}}{\sum_j \mathrm{Y}_j(t)\mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_j(t)}} \right\}^{\triangle\overline{\mathrm{N}}_i(t)}.$$

A função log-verosimilhança parcial é,

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{t \geq 0} \left[\mathrm{Y}_i(t)\boldsymbol{\beta}'\mathbf{z}_i(t) - \ln\left(\sum_j \mathrm{Y}_i(t)\mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_j(t)} \right) \right] \triangle\,\overline{\mathrm{N}}_i(t).$$

Sejam,

$$\mathbf{S}^{(0)}(\boldsymbol{\beta}, t) = \textstyle\sum_j \mathrm{Y}_j(t)\mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_j(t)};$$
$$\mathbf{S}^{(1)}(\boldsymbol{\beta}, t) = \textstyle\sum_j \mathrm{Y}_j(t)\mathbf{z}_j(t)\mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_j(t)};$$
$$\mathbf{S}^{(2)}(\boldsymbol{\beta}, t) = \textstyle\sum_j \mathrm{Y}_j(t)\mathbf{z}_j(t)\mathbf{z}_j(t)'\mathrm{e}^{\boldsymbol{\beta}'\mathbf{z}_j(t)}.$$

O estimador de máxima verosimilhança parcial de $\boldsymbol{\beta}$ é obtido resolvendo o sistema de equações $\boldsymbol{\mu}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$ em que,

$$\mu(\boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{t \geq 0} \left[\mathbf{z}_i(t) - \frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, t)}{\mathbf{S}^{(0)}(\boldsymbol{\beta}, t)} \right] \triangle \overline{\mathrm{N}}_i(t).$$

A matriz de informação de Fisher correspondente é,

$$\mathbf{I}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{t \geq 0} \left[\frac{\mathbf{S}^{(2)}(\boldsymbol{\beta}, t)}{\mathbf{S}^{(0)}(\boldsymbol{\beta}, t)} - \frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, t)\mathbf{S}^{(1)}(\boldsymbol{\beta}, t)'}{\mathbf{S}^{(0)}(\boldsymbol{\beta}, t)^2} \right] \triangle \overline{\mathrm{N}}_i(t).$$

A notação anterior pode ser usada tanto no caso em que se considera apenas o tempo até ao primeiro acontecimento, como no caso em que se considera a ocorrência de acontecimentos múltiplos.

Apêndice C

Glossário de alguns termos usados em Cardiologia

Acidente vascular cerebral/Acidente isquémico transitório (AVC/AIT)

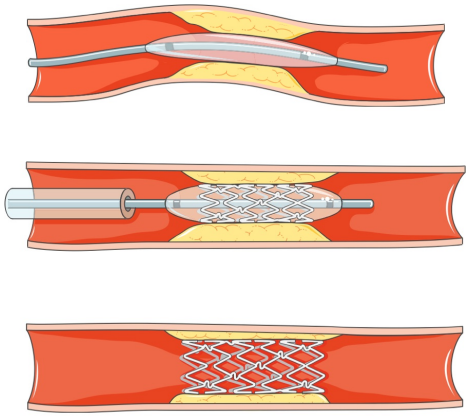
Um AVC ou "trombose"é a consequência do entupimento súbito de uma artéria cerebral. As células da parte do cérebro que essa artéria alimenta começam rapidamente a morrer e a parte do corpo que essa região cerebral controla deixa de funcionar. O AIT popularmente chamado “princípio de AVC”, surge quando o entupimento da artéria cerebral é transitório e os sintomas desaparecem espontaneamente em menos de 24 horas.

Angina de peito

A angina de peito é uma dor breve resultante da falta temporária de oxigénio no músculo cardíaco por redução do fluxo de sangue, devido à obstrução das artérias coronárias por placas de gordura.

Angioplastia coronária (PTCA)

A angioplastia é um método mecânico que usa a passagem de um cateter para repor a circulação coronária. Com este processo, é possível passar um balão ou um *stent* através da área de estreitamento que, ao ser insuflado na zona ocluída, aumenta o diâmetro da artéria e restabelece a circulação sanguínea.



Antihipertensor

É uma terapêutica utilizada para controlar a hipertensão arterial quando aliada a um estilo de vida saudável. Este tipo de medicamentos serve para reduzir o risco de o doente vir a sofrer problemas cardiovasculares mais graves que podem, inclusive ser mortais. A duração do tratamento depende da situação clínica do doente.

Antiplaquetário

É uma terapêutica utilizada para impedir a formação de coágulos na corrente sanguínea. Dois exemplos são a aspirina e o clopidogrel (muitas vezes usados em associação). Nos doentes que tiveram um EAM, que têm uma angina de peito ou que foram submetidos a PTCA ou a CABG, a toma diária de aspirina reduz de forma substancial o risco de um segundo EAM e de vir a sofrer um AVC.

Arritmia

É uma perturbação do ritmo dos batimentos cardíacos que pode ter consequências fatais quando não tratada. De acordo com a frequência cardíaca, as arritmias classificam-se em dois grupos: as taquiarritmias, quando se ultrapassam os 100 batimentos por minuto; e as bradiarritmias, quando se produzem menos de 60 batimentos por minuto. O tipo de arritmia pode ser identificado através de um eletrocardiograma.

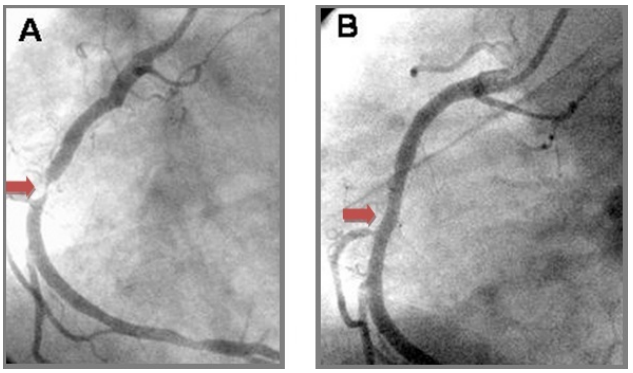
Betabloqueante (BB)

É uma terapêutica utilizada no tratamento de diversas doenças cardíacas em que é necessário reduzir a necessidade de consumo de oxigénio pelo miocárdio. Este fármaco diminui a frequência dos batimentos cardíacos, a pressão arterial e a contractilidade. Nos doentes que sofreram um EAM a toma continuada deste fármaco reduz o risco de arritmias, de morte e de um novo EAM.

Biomarcadores cardíacos

São substâncias presentes dentro das células do coração necessárias ao seu normal funcionamento. Quando os seus níveis no sangue são elevados, pode ser um sinal de lesão das células cardíacas, por exemplo, no EAM. Os biomarcadores mais utilizados são a troponina e a CK-MB.

Cateterismo cardíaco



O cateterismo cardíaco, também designado por angiografia coronária, consiste na passagem de um tubo fino e flexível (cateter) através de uma artéria do braço ou da virilha, com o doente acordado e usando apenas anestesia local. Este cateter vai progredir ao longo da artéria picada até ao local das artérias coronárias, as quais, após a injeção de um produto de contraste para os raios X, são visualizadas, podendo analisar-se a localização e a gravidade das obstruções. Na figura A está representada uma obstrução grave numa

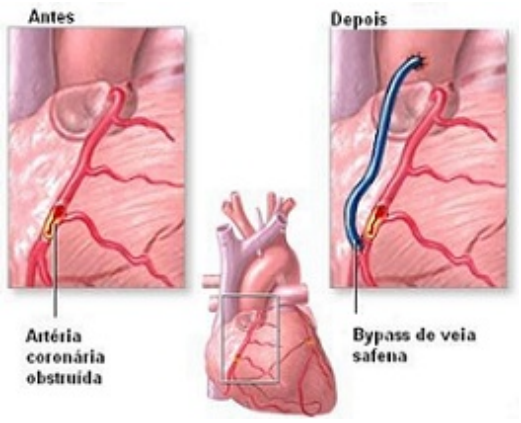
artéria, enquanto a figura B apresenta a mesma artéria, já depois do restabelecimento da circulação sanguínea.

Choque cardiogénico

O choque cardiogénico é um estado no qual o coração, subitamente enfraquecido, não é capaz de bombear sangue suficiente para as necessidades do organismo. A causa mais comum de choque é o EAM. Por sua vez, após um EAM, o choque é a causa mais frequente de morte.

Cirurgia de *bypass* (CABG)

Esta cirurgia permite melhorar e restabelecer o fluxo sanguíneo do coração através da utilização de uma artéria saudável (artéria mamária interna, veia safena ou artéria radial). Estas veias e artérias permitem estabelecer uma ponte, contornando o local de obstrução da artéria coronária. A figura seguinte mostra um exemplo.



Classe killip (Classe kk)

A classe Killip-Kimball ou classe KK é uma escala de classificação do grau de insuficiência cardíaca em indivíduos que sofreram EAM. Os doentes podem ser classificados segundo quatro classes: Classe I, sem evidência clínica de insuficiência cardíaca; Classe II, insuficiência cardíaca ligeira; Classe III, insuficiência cardíaca grave ou presença de edema pulmonar; e Classe IV, presença de choque cardiogénico caracterizado por hipotensão (TAS $<$ 90 mmHg) e evidência de vasoconstrição periférica. Quanto maior a classe KK pior é o prognóstico do doente.

Colesterol/ Colesterol LDL/ Colesterol HDL

O colesterol é uma gordura que circula no sangue, que em pequenas quantidades, é essencial ao normal funcionamento das células do organismo. A maior parte do colesterol é produzida no fígado, sendo o restante absorvido ao nível do intestino a partir dos alimentos que ingerimos. O colesterol LDL ("mau colesterol"), promove a acumulação do colesterol em vários órgãos, tecidos e parede dos vasos arteriais. Por seu lado, o colesterol HDL ("bom colesterol") remove o excesso de colesterol das células e promove o seu transporte para o fígado.

Creatinina

A creatinina é um ótimo parâmetro de avaliação da função renal, já que a sua produção apenas depende do metabolismo celular muscular e é quase exclusivamente eliminada por filtração glomerular. A sua eliminação é exclusivamente urinária.

Diabetes *mellitus* (DM)

A maioria dos alimentos que comemos é convertida pelo nosso organismo em glicose. A glicose é um tipo de açúcar que depois de absorvida pelas células do organismo serve de fonte de energia. Na presença de diabetes o nosso organismo ou não produz insulina suficiente (DM tipo 1) ou não utiliza a sua própria insulina tão bem como devia (DM tipo 2), o que leva à presença de glicémia (níveis de açúcar no sangue) elevada. A diabetes é diagnosticada quando a glicémia em jejum é superior ou igual a 126 mg/dL.

Dislipidémia

Manifesta-se quando os valores do colesterol no sangue são superiores aos níveis máximos recomendados em função do risco cardiovascular individual.

Doença arterial periférica (DAP)

A doença arterial periférica caracteriza-se por uma diminuição do fluxo sanguíneo nos membros superiores e inferiores devido a obstruções das artérias, dificultando a oxigenação dos músculos.

Eletrocardiograma (ECG)

É um exame no qual é feito o registo da variação dos potenciais elétricos gerados pela atividade elétrica do coração. Este exame é utilizado no diagnóstico de doenças cardíacas, em especial as arritmias cardíacas e o EAM.

Estatina

É um medicamento que interfere principalmente na produção de colesterol ao nível do fígado. Atua principalmente à custa da diminuição dos níveis do LDL. O seu benefício vai para além do conseguido através da redução dos valores de colesterol, evitando a progressão da aterosclerose.

Frequência cardíaca (FC)

Frequência cardíaca é determinada pelo número de batimentos cardíacos por unidade de tempo, geralmente expressa em batimentos por minuto (bpm). A FC pode variar de acordo com a necessidade de

oxigénio do organismo. Durante o exercício físico a FC eleva-se devido a uma elevada necessidade de oxigénio, já durante o sono o seu valor é mais baixo.

Hemorragia

A hemorragia consiste numa perda de sangue do sistema circulatório. As hemorragias são classificadas em quatro classes, de acordo com o volume de sangue perdido: classe 1, quando se perde até 15% de volume; classe 2, quando o volume da perda varia entre 15% e 30%; classe 3, quando o volume da perda varia entre 30% e 40%; classe 4, quando o volume de perda é superior a 40%.

Hipertensão (HTA)

A tensão arterial considera-se elevada quando TAS $\geq$ 140 mmHg ou TAD $\geq$ 90 mmHg. A HTA é um fator de risco para as doenças cardiovasculares, uma vez que as artérias sujeitas a uma tensão excessiva tornam-se mais espessas e rígidas, o que favorece a progressão da aterosclerose.

Índice de Massa Corporal (IMC)

Indicador utilizado para avaliar a relação entre o peso e a estatura. Calcula-se dividindo o peso (em kg) pelo quadrado da estatura (em m^2).

Inibidor da enzima da conversão da angiotensina (IECA)

Salienta-se o seu papel na constrição dos vasos sanguíneos e no consequente aumento da pressão arterial, que provoca um aumento do trabalho do coração para bombear o sangue. Contribui ainda para o aumento da espessura do músculo do coração e da parede das artérias coronárias. Após um EAM a toma destes fármacos associa-se a uma melhoria do prognóstico.

Insuficiência cardíaca (ICC)

Considera-se que existe insuficiência cardíaca quando, por doença, o coração deixa de ser capaz de cumprir com eficiência a sua função de bombear o sangue para fora do coração, ficando este acumulado nos pulmões, o que leva à insuficiência respiratória. Duas das causas mais frequentes de insuficiência cardíaca são a HTA e o EAM.

Insulina

Hormona produzida pelo pâncreas cujos níveis, em circunstâncias normais, estão estreitamente relacionados com os níveis de açúcar no sangue. A elevação da Glicemia (açúcar no sangue) é um estímulo para a produção de insulina. Esta promove a absorção da Glicose pelas células, permitindo o seu

aproveitamento como fonte de energia. A insulina é também responsável pela acumulação de Glicose no fígado sob a forma de glicogénio, que funciona como fonte de energia quando é reconvertido em Glicose em situações de stress ou exercício.

***Stent* (coronário)**

É uma espécie de "mola" de metal utilizada na angioplastia coronária. O *stent* é posicionado na altura do entupimento e é expandido para ficar colado à parede interna do vaso sanguíneo.

Tabagismo

É um importante fator de risco para doenças pulmonares e cardiovasculares graves. O fumo do tabaco contém mais de 4000 substâncias químicas, várias das quais com efeitos tóxicos, irritantes ou cancerígenos. A nicotina aumenta a tensão arterial, a frequência cardíaca, diminui o débito cardíaco e o fluxo de sangue nas artérias coronárias. O tabaco torna os vasos rígidos e promove a formação de coágulos, favorece o depósito de colesterol que resulta em aterosclerose e trombose aguda.

Taxa de filtração glomerular (TFG)

Volume de água filtrada fora do plasma pelas paredes dos capilares glomerulares nas cápsulas de Bowman, por unidade de tempo. A filtração glomerular é a primeira etapa na formação da urina. O sangue arterial é conduzido sob alta pressão nos capilares do glomérulo. Essa pressão tem intensidade suficiente para que parte do plasma passe para a cápsula de Bowman, onde as substâncias pequenas (água, sais, vitaminas, açúcares e aminoácidos) saem do glomérulo e entram na cápsula de Bowman.

Tensão arterial (TA)/ Tensão arterial sistólica (TAS)/ Tensão arterial diastólica (TAD)

A tensão arterial é a pressão do sangue dentro do coração e das artérias. É descrita por dois valores, tensão arterial sistólica e tensão arterial diastólica, vulgarmente conhecidas como tensão "máxima" e "mínima" respetivamente. A TAS mede a pressão provocada pela contração do coração, enquanto a TAD quantifica a pressão nas artérias quando o coração relaxa entre duas contrações.

Triglicerídeos (TG)

Os triglicéridos podem ter uma origem endógena, quando produzidos no fígado ou no tecido adiposo, principalmente a partir de hidratos de carbono; ou exógena, quando são provenientes da alimentação. São armazenados no tecido adiposo e muscular como fonte de energia e, de acordo com as necessidades energéticas do organismo vão sendo regularmente libertados e hidrolisados em glicerol e ácidos gordos. O valor de TG juntamente com o valor de CT, HDL e LDL caracterizam o perfil lipídio.

Trombólise

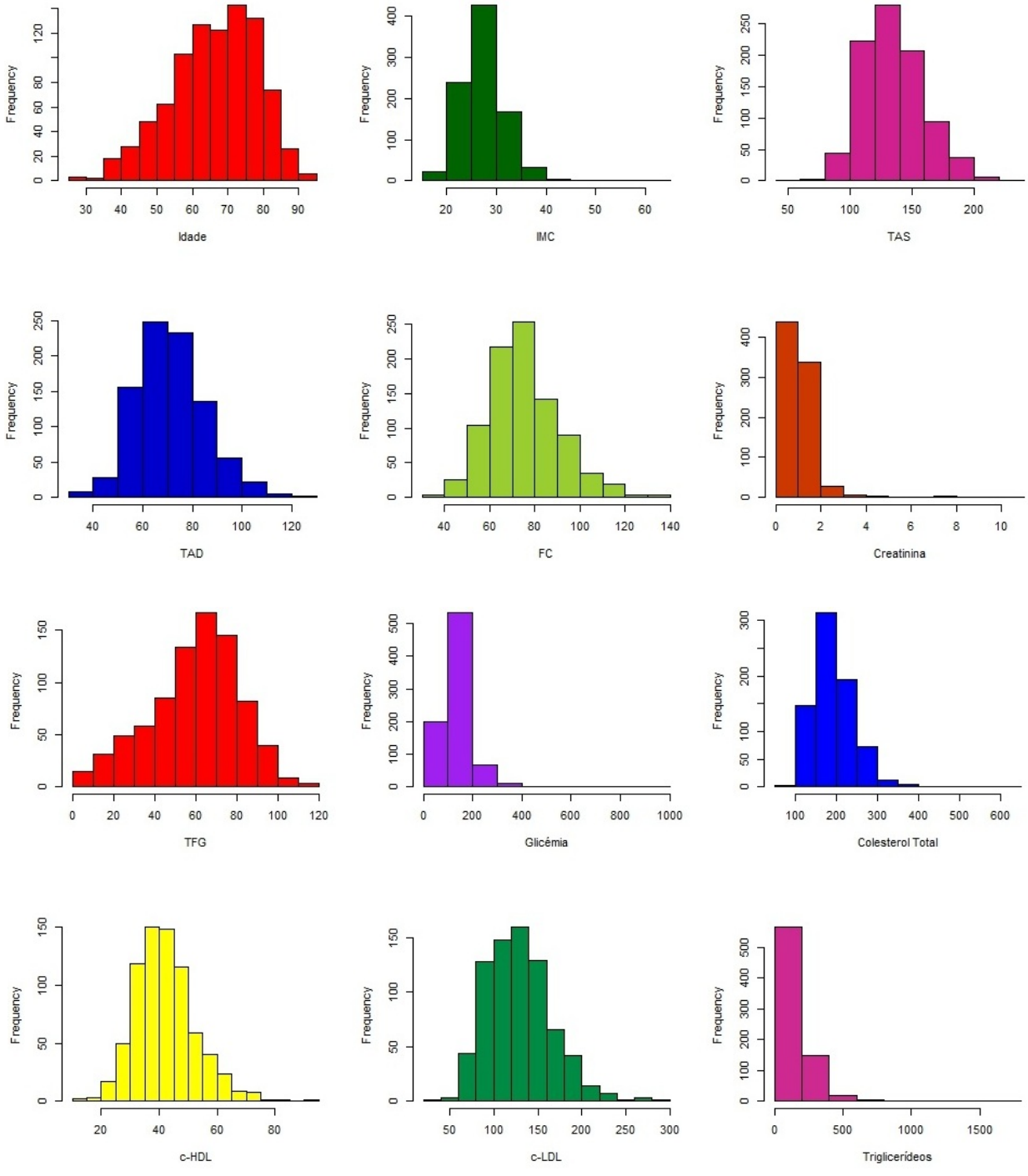
A trombólise consiste na administração intravenosa de um medicamento chamado trombolítico que tem como função dissolver um coágulo. Foi o primeiro tratamento eficaz e mudou radicalmente o prognóstico sombrio da fase inicial do EAM.

Apêndice D

Gráficos

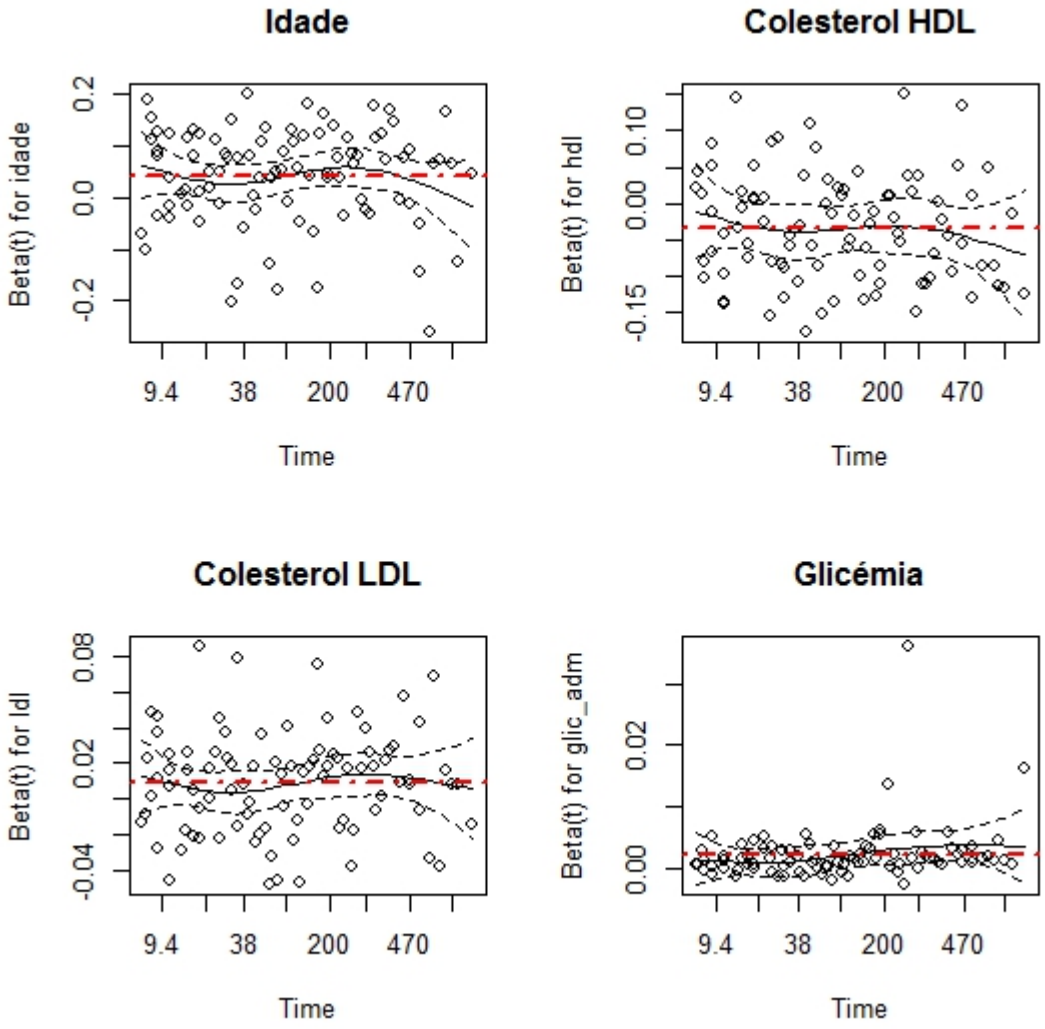
D.1 Histogramas

Histogramas da distribuição dos valores das covariáveis contínuas.



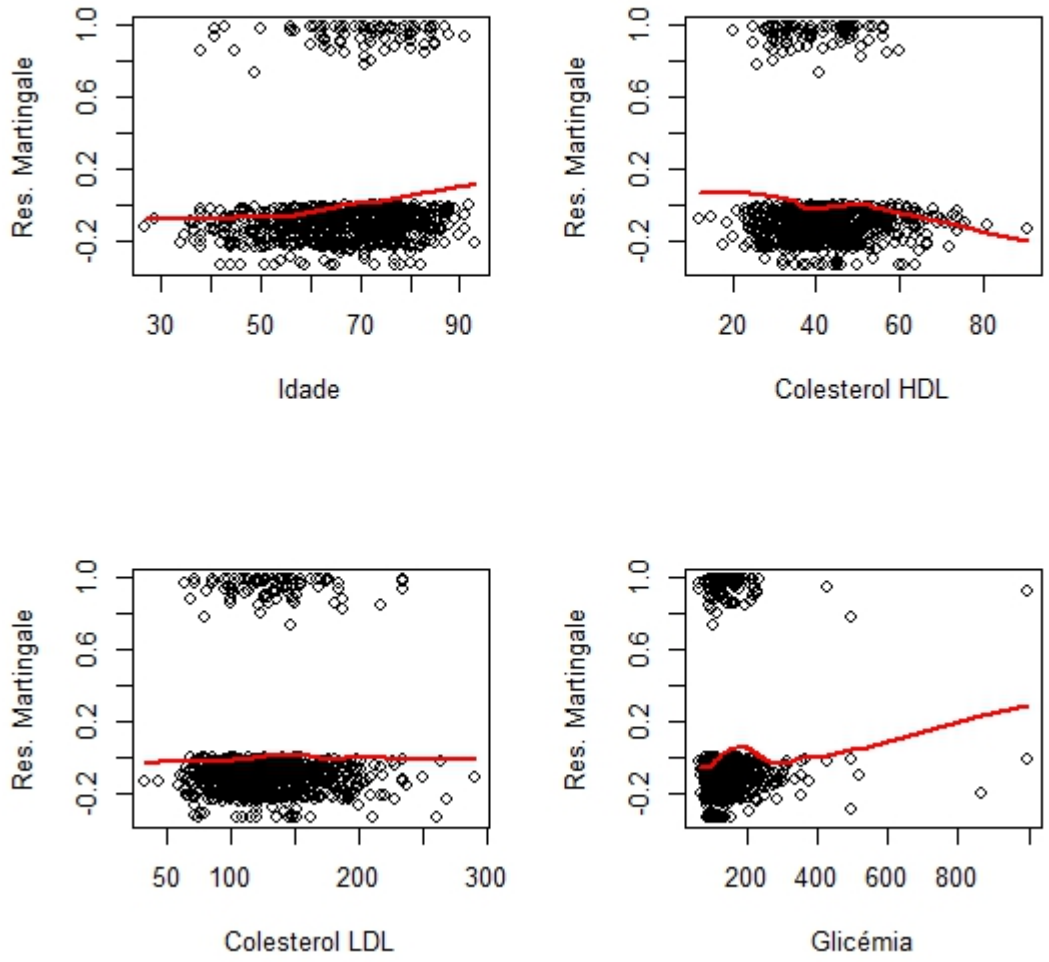
D.2 Gráficos dos resíduos de Schoenfeld ponderados

Gráficos dos resíduos de Schoenfeld ponderados para o primeiro modelo PWP-GT.



D.3 Gráficos dos resíduos martingala

Gráficos dos resíduos martingala para o primeiro modelo PWP-GT.



Bibliografia

Anderson, P. K., Borgan, O., Gill, R. D. e Keiding, Linear nonparametric tests for comparison of counting processes, with applications to censored survival data. *Internat. Statist. Rev.*, 50 (3):219-258, 1982.

Anderson, P. K., Borgan, O., Gill, R. D. e Keiding, N. *Statistical models based on counting processes.* Springer-Verlag. New York. 1993.

Anderson, P. K. e Gill, R. D. Cox's regression model for counting processes: a large sample study. *Annals of Statistics*, 10 (4):1100-1120, 1982.

Barlow, W. E. e Prentice, R.L. Residuals for relative risk regression. *Biometrika*, 75 (1):65-74, 1988.

Box-Steffensmeier, J. e De Boef, S. Repeated events survival models: the conditional frailty model. *Statistics in Medicine*, 25 (20): 3518-3533, 2006.

Breslow, N. E. Contribution to discussion of paper by D.R. Cox. *Journal of Royal Statistical Society*: Series B, 34 (2):216-217, 1972.

Breslow, N. E. Covariance analysis of censored survival data. *Biometrics*, 30 (1):89-99, 1974.

Cai, J. e Schaubel, D. E. Analysis of recurrent event data. Em *Handbook of Statistics, vol. 23.* Elsevier B. V. 2004.

Cox, D. R. Regression models and life tabels (with discussion). *Journal of Royal Statistical Society*: Series B, 34 (2):187-220, 1972.

Cox, D. R. Partial likelihood. *Biometrika*, 62 (2):269-276, 1975.

Efron, B. The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association*, 72 (359):557-565, 1977.

Fleming, T. R. e Harrington, D. P. Nonparametric estimation of the survival distribution in censored data. *Comm. Stat. Theory Methods*, 13 (20):2469-2486, 1984.

Fleming, T. R. e Harrington, D. P. *Counting processes and survival analysis.* John Wiley & Sons, Inc. New York. 1991.

Gavina, C., Pinho, T. *Enfarte agudo do miocárdio.* Disponível em: http://www.spc.pt/, 2010.

Gehan, E. A. A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika*, 52 (1-2):203-223, 1965.

Grambsch, P. M. e Therneau, T. M. Proportional hazards tests in diagnostics based on weighted residuals. *Biometrika*, 81 (3):515-526, 1994.

Hosmer, D. W. e Lemeshow, S. *Applied survival analysis: regression modeling of time to event data.* John Wiley & Sons, Inc. New York. 1999.

Kalbfleisch, J. D. e Prentice, R. L. *The statistical analysis of failure time data.* John Wiley & Sons, Inc. New York. 1980.

Kaplan, E. L. e Meier, P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53 (282):457-481, 1958.

Kelly, P. J. e Lim, L. L-Y. Survival analysis for recurrent event data: an application to childhood infectious diseases. *Statistics in Medicine*, 19 (1): 13-33, 2000.

Lee, E. W., Wei, L. J. e Amato, D. A. Cox-type regression analysis for large numbers of small groups of correlated failure time observations. Em Klein, J. P. e Goel, P. K. (eds), *Survival Analysis: State of the Art*, Kluwer Academic Publisher, Dordrecht, pp. 237-247, 1992.

Lin, D. Y. Cox regression analysis of multivariate failure time data: the marginal approach. *Statistics in Medicine*, 13 (21): 2233-2247, 1994.

Marubini, E. e Valsecchi, M. G. *Analysing survival data from clinical trials and observational studies.* John Wiley & Sons, Ltd. Chichester. 1995.

Peto, R. Contribution to discussion of paper by D.R. Cox. *Journal of Royal Statistical Society*: Series B, 34 (2):205-207, 1972.

Prentice, R. L., Williams, J. e Peterson, A. V. On the regression analysis of multivariate failure time data. *Biometrika*, 68 (2):373-379, 1981.

Schoenfeld, D. Partial residuals for the proportional hazards regression model. *Biometrika*, 69 (1):239-241, 1982.

Tarone, R. E. e Ware J. On distribution-free tests for equality of survival distributions. *Biometrika*, 64 (1):156-160, 1977.

Therneau, T. M., Grambsch, P. M. e Fleming T.R. Martingale-based residuals for survival models. *Biometrika*, 77 (1):147-160, 1990.

Therneau, T. M. e Grambsch, P. M. *Modeling survival data - extending the Cox model.* Springer-Verlag, New York. 2000.

Wei, L. J., Lin, D. Y., Weissfeld, L. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association*, 84 (408):1065-1073, 1989.